Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition

Romain Serizel, Diego Giuliani

VOCAL TRACT LENGTH NORMALISATION APPROACHES TO DNN-BASED CHILDREN'S AND ADULTS' SPEECH RECOGNITION

Romain Serizel and Diego Giuliani

HLT research unit, Fondazione Bruno Kessler (FBK), Trento, Italy
(serizel,giuliani)@fbk.eu

ABSTRACT

This paper introduces approaches based on vocal tract length normalisation (VTLN) techniques for hybrid deep neural network (DNN) - hidden Markov model (HMM) automatic speech recognition when targeting children's and adults' speech. VTLN is investigated by training a DNN-HMM system by using first mel frequency cepstral coefficients (MFCCs) normalised with standard VTLN. Then, MFCCs derived acoustic features are combined with the VTLN warping factors to obtain an augmented set of features as input to a DNN. In this later, novel, approach the warping factors are obtained with a separate DNN and the decoding can be operated in a single pass when standard VTLN approach requires two decoding passes. Both VTLN-based approaches are shown to improve phone error rate performance, up to 20% relative improvement, compared to a baseline trained on a mixture of children's and adults' speech.

Index Terms— Vocal tract length normalisation, automatic speech recognition, children's speech recognition, deep neural networks

1. INTRODUCTION

Speaker-related acoustic variability is one of the major source of errors in automatic speech recognition. In this paper we cope with age group differences, by considering the relevant case of children versus adults, as well as with male/female differences. Here vocal tract length normalisation (VTLN) is used together with deep neural network (DNN) to deal with the acoustic variability induced by age and gender differences.

Developmental changes in speech production introduce age-dependent spectral and temporal variabilities in speech produced by children. Studies on morphology and development of the vocal tract [1] reveal that during childhood there is a steady gradual lengthening of the vocal tract as the child grows while a concomitant decrease in formant frequencies occurs [2, 3]. In particular, for females there is an essential gradual continuous growth of vocal tract through puberty into adulthood, while for males during puberty there is a disproportionate growth of the vocal tract, which lowers formant frequencies, together with an enlargement of the glottis, which lowers the pitch. After age 15, males show a substantially longer vocal tract and lower formant frequencies than females. As consequence, voices of children tend to be more similar to the voices of women than to those of men. For adults, variations in voice characteristics due to speaker age are much less evident than for children while males and females exhibit clearly different formant patterns.

VTLN aims at reducing inter-speaker acoustic variability by warping the frequency axis of the speech power spectrum to account for the fact that the precise locations of vocal-tract resonances vary roughly monotonically with the physical size of the speaker [4, 5, 6]. Effectiveness of VTLN techniques was widely proven in the past for hidden Markov model (HMM) - Gaussian mixture modelisation (GMM) based recognition of children's and adults' speech [7, 8, 4, 5, 6].

During the past years, DNN has proven to be an effective alternative to HMM-GMM based ASR [9, 10] obtaining good performance with context dependent hybrid DNN-HMM [11, 12]. Very recently the use of DNN has been also investigated for ASR of children's speech [13].

In [14] an investigation was conducted by training a DNN on VTLN normalised acoustic features, it was found that in a large vocabulary adults' speech recognition task limited gain can be achieved with respect to using unnormalised acoustic features. It was argued that, when a sufficient amount of training data is available, DNN are already able to learn, to some extent, internal representations that are invariant with respect to sources of variability such as the vocal tract length and shape. However, when only limited training data is available from a heterogeneous population of speakers, made of children and adults as in our case, the DNN might not be able reach strong generalisation capabilities [15]. In such case, techniques like DNN adaptation [16, 17, 18], speaker adaptation [19, 20] or VTLN [4, 5, 6] can help to improve the performance. Here we consider first the application of a conventional VTLN technique to normalise MFCC as input features to a DNN-HMM.

Recent works have shown that augmenting the inputs of a DNN with, e.g. an estimate of the background noise [21]

This work was partially funded by the European project EU-BRIDGE, under the contract FP7-287658.

or utterance i-vector [22], can improve the robustness and speaker independence of the DNN. We then propose to augment the MFCC inputs of the DNN with the posterior probabilities of the VTLN-warping factors to improve robustness with respect to inter-speaker acoustic variations.

The rest of this paper is organized as follows. Section 2 introduces the general DNN baseline and reminds DNN adaptation as a contrastive way to deal with acoustic variability by targeting groups of speakers. Section 3 introduces two approaches, based on VTLN, to cope with inter-speaker acoustic variability. Experimental setup is described in Section 4 and results are presented in Section 5. Finally, conclusions are provided in Section 6.

2. DNN-HMM BASELINES

Performance obtained with the VTLN techniques are to be confronted with results obtained with a general DNN-HMM baseline system and with those obtained in our previous work [15] using an approach based on a DNN adaptation procedure similar to the procedure proposed in [18] for the case of multilingual training.

2.1. General DNN-HMM

The DNN-HMM baseline is trained on speech collected from speakers from all target groups, that is in our case children, adult males and adult females. This training procedure is an attempt to achieve a DNN-HMM system with strong generalisation capabilities.

2.2. Age/gender specific DNN-HMM

Estimating the DNN parameters on speech from all groups of speakers, may however, have some limitation due to the inhomogeneity of the speech data that may negatively impact on the classification accuracy compared to group-specific DNN. One option is to train group specific DNN. However, in our case only limited data is available for each specific group of speakers so that the DNN might not be able to be properly trained. To overcome this problem, the DNN trained on all data available is adapted to each specific group of speakers by using group specific training data. Further details on this approach can be found in [15].

3. VTLN FOR DNN-HMM

In this work, the problem of inter-speaker acoustic variability due to vocal tract length (and shape) variations among speakers is tackled with two different approaches. The first one is based on the conventional VTLN approach [4, 5, 6]. The resulting VTLN normalised acoustic features are used as input to the DNN both during training and testing [14]. The second approach, proposed in this paper, has two main features: a) by using a dedicated DNN, for each speech frame the posterior probability of each warping factor is estimated and b) for each speech frame the vector of the estimated warping factor posterior probabilities is appended to the unnormalised acoustic features vector, extended with context, to form an augmented acoustic features vector for the DNN-HMM system.

3.1. VTLN normalised features as input to the DNN

In the conventional frequency warping approach to speaker normalisation [4, 5, 6], typical issues are the estimation of a proper frequency scaling factor for each speaker, or utterance, and the implementation of the frequency scaling during speech analysis. A well known method for estimating the scaling factor is based on a grid search over a discrete set of possible scaling factors by maximizing the likelihood of warped data given a current set of HMM-based acoustic models [4]. Frequency scaling is performed by warping the power spectrum during signal analysis or, for filter-bank based acoustic front-end, by changing the spacing and width of the filters while maintaining the spectrum unchanged [4]. In this work we adopted the latter approach. Details on the VTLN implementation are provided in Section 4.3.

Similarly to as proposed in [14], the VTLN normalised acoustic features are used to form the input to the DNN-HMM system both during training and testing.

3.2. Posterior probabilities of VTLN warping factors as input to DNN

In this approach we propose to augment the acoustic features vector with the posterior probabilities of the VTLN warping factors to train a warping-factor aware DNN. Similar approaches have recently been shown to improve the robustness and speaker independence of the DNN [21, 22].

The VTLN procedure is first applied to generate a warping factor for each utterance in the training set. Then, training utterances and corresponding warping factors are used to train a DNN classifier that learns to infer the VTLN warping factor from the acoustic feature vector. This DNN is then used to produce the posterior probabilities of the VTLN warping factors for each input speech frame. This DNN will be referred to as DNN-warp.

During training and testing of the DNN-HMM system, for each speech frame the warping factor posterior probabilities are estimated with the DNN-warp. These estimated posterior probabilities are appended to the unnormalised acoustic features vector, extended with context, to form an augmented acoustic features vector. The extended features vector is then normalised and used as input to the DNN-HMM.

This approach has the advantage to reduce considerably the complexity during decoding compared to the approach making use of VTLN normalised acoustic features that requires two decoding passes [4, 23].

4. EXPERIMENTAL SETUP

4.1. Speech corpora

For this study we relied on two Italian speech copora: the ChildIt corpus consisting of children speech and the APASCI corpus consisting of adults' speech. Both corpora were used for evaluation purposes, while the ChildIt and the APASCI provided similar amount of training data for children and adults, respectively.

4.1.1. ChildIt

The ChildIt corpus [8, 24] is an Italian, task-independent, speech corpus that consists of clean read speech from children aged from 7 to 13 years, with a mean age of 10 years. Children in the ChildIt corpus were evenly distributed by grade, from grade 2 through grade 8. Children in grade 2 were approximately 7 years old while children in grade 8 were approximately 13 years old. The overall duration of audio recordings in the corpus is 10h:48m. Speech was collected from 171 children, each child read 58 or 65 sentences selected from electronic texts concerning literature for children, depending on his/her grade. Each speaker read a different set of sentences which included, however, a set of phonetically rich sentences (5-8 sentences) which were repeated by several speakers. Speech was acquired at 16 kHz, with 16 bit accuracy, using a Shure SM10A head-worn microphone. The corpus was partitioned into: a training set consisting of data from 115 speakers for a total duration of 7h:15m; a development set consisting of data from 14 speakers (1 male and 1 female per grade), for a total durations of 0h:49m (24,880 phone occurrences); a test set consisting of data from 42 speakers balanced with respect to age and gender (that is 3 males and 3 females per grade) for a total duration of of 2h:20m (74,596 phone occurrences). Repetitions of phonetically rich sentences were not included in the development and test sets. The development set is formed by 767 audio recordings while the test set is formed by 2299 audio recordings. 74 audio recordings in the development set correspond to repetitions of sentences also appearing in the training set, similarly 128 audio recordings in the test set correspond to repetitions of sentences also appearing in the training set.

4.1.2. APASCI

The APASCI speech corpus [25] is a task-independent, high quality, acoustic-phonetic Italian corpus. APASCI was developed at ITC-irst and consists of speech data collected from 194 adult speakers for a total durations of 7h:05m. Acquisitions were performed in quiet rooms using a digital audio tape recorder and a high quality close talk microphone. Audio signals were down-sampled from 48 kHz to 16 kHz with 16 bit accuracy. Most of the speaker performed a single recording session, while 44 speakers performed two recording sessions. In each recording sessions each speaker read a calibration sentence, 4 phonetically reach sentences and 15 or 20 diphonically rich sentences.

The corpus was partitioned into: a training set consisting of data from 134 speakers for a total duration of 5h:19m; a development set consisting of data from 30 speakers balanced per gender, for a total durations of 0h:39m (20,363 phone occurrences); a test set consisting of data from 30 speakers balanced per gender, for a total duration of 0h:40m (20,708 phone occurrences). Audio recordings of phonetically rich sentences and of the calibration sentence were not included in the development and test sets. The development set is formed by 550 audio recordings while the test set is formed by 520 audio recordings. 254 audio recordings in the development set correspond to repetitions of sentences also appearing in the training set, furthermore 170 audio recordings in the test set correspond to repetitions of sentences also appearing in the training set.

4.2. ASR systems

4.2.1. General DNN-HMM

The acoustic features are 13 MFCC, including the zero order coefficient, computed on 20ms frames with 10ms overlap. The context spans on a 31 frame window on which Hamming windowing is applied. This 403 dimensional features vector is then projected to a 208 dimensional features vector by applying Discrete Cosine Transform (DCT) and normalised before being used as input to the DNN. The targets of the DNN are the 3039 tied-states obtained from triphone HMM-GMM models based on a set of 48 phonetic units derived from the SAMPA Italian alphabet and trained on the mixture of adults' and children's speech (ChildIt + APASCI). The DNN has 4 hidden layers, each of which contains 1500 elements such that the DNN architecture can be summarised as follows: 208 x 1500 x 1500 x 1500 x 1500 x 3039.

The DNN are trained with the TNet software package [26]. The DNN weights are initialised randomly and pre-trained with restricted Boltzmann machines (RBM) [27, 28]. The first layer is pre-trained with a Gaussian-Bernouilli RBM trained during 10 iterations with a learning rate of 0.005. The following layers are pre-trained with a Bernouilli-Bernouilli RBM trained during 5 iterations with a learning rate of 0.05. Mini-batch size is 250. For the back propagation training the learning rate is kept to 0.02 as long as the frame accuracy on the cross-validation set progresses by at least 0.5% between successive epochs. The learning rate is then halved at each epoch until the frame accuracy on the cross-validation set fails to improve by at least 0.1%. The mini-batch size is 512. In both pre-training and training, a first-order momentum of 0.5 is applied.

4.2.2. Age/gender specific DNN-HMM

The DNN-HMM described above is adapted to each of the three target groups of speakers by using the available training data as in [15] to obtain three group specific DNN-HMM systems. At recognition time, each utterance is decoded with the matching group specific DNN-HMM system. Note, however, that to operate fully automatically this procedure would require a classifier to perform the selection of the appropriate DNN-HMM system.

4.3. VTLN

In this work we are considering a set of 25 warping factors evenly distributed, with step 0.02, in the range 0.76-1.24. During both training and testing a grid search over the 25 warping factors was performed. The acoustic models for scaling factor selection, carried out on an utterance-by-utterance basis, were speaker-independent triphone HMM with 1 Gaussian per state and trained on unwarped children's and adults' speech [23, 24].

The DNN-warp inputs are the MFCC with a 61 frames context window, DCT projected to a 208 dimensional features vector. The targets are the 25 warping factors. The DNN has 4 hidden layers, each of which contains 500 elements such that the DNN architecture can be summarised as follows: 208 x 500 x 500 x 500 x 500 x 25. The training procedure is the same as for the DNN acoustic model in the DNN-HMM.

The posterior probabilities obtained with the DNN-warp are concatenated with the 208-dimensionnal DCT projected acoustic features vector to produce a 233-dimensional features vector that is normalised before being used as input to the DNN. The new DNN acoustic model has 4 hidden layers, each of which contains 1500 elements such that the DNN architecture can then be summarized as follows: 233 x 1500 x 1500 x 1500 x 1500 x 3039.

5. EXPERIMENT RESULTS

The experiments presented here are designed to verify the following hypothesis:

- VTLN can be beneficial to the DNN-HMM framework when targeting a heterogeneous speaker population with limited amount of training data
- Developing an "all-DNN" approach to VTLN for DNN-HMM framework, when targeting a heterogeneous speaker population, offers a credible alternative to the use of VTLN normalised acoustic features or to the use of age/gender group specific DNN.

During the experiments the language model weight is tuned on the development set and used to decode the test set. Results were achieved with a phone loop language model and the phone error rate (PER) was computed based on a reduced set of 28 phone labels. Variations in recognition performance were validated using the matched-pair sentence test [29] to ascertain whether the observed results were inconsistent with the null hypothesis that the output of two systems were statistically identical. Considered significance levels were $.05$, $.01$ and $.001$.

5.1. Phone error rate performance

Table 1 presents the PER obtained with the DNN-HMM baseline, and the VTLN approaches: the VTLN applied to MFCC during training and testing (row *VTLN-normalisation*) and the MFCC features vector augmented with the posterior probabilities of the warping factors (row *VLTN + MFCC*). The results are compared with PER obtained when using a matching adapted DNN-HMM (row *DNN-adaptation*). On the evaluation set including all the target groups of speakers (ChildIt + APASCI) the VTLN normalisation approach improve the baseline performance by 18% relative (from 14.29% to 12.00% PER with $p < .001$) whereas the system working with the MFCC features vector augmented with the posterior probabilities of the warping factors improves the baseline by 9% relative (from 14.29% to 13.12% PER with $p < .001$). The performance difference between VTLN and DNN-adaptation (from 12.00% to 11.59% PER) is not statistically significant.

For children speakers, the VTLN normalisation approach improve the baseline performance by 22% relative (from 15.56% to 12.80% PER with $p < .001$) and the system working with the MFCC features vector augmented with the posterior probabilities of the warping factors improves the baseline by 10% relative (from 15.56% to 14.10% PER with $p < .001$). The performance difference between VTLN and DNN-adaptation (from 12.80% to 12.43% PER) is not statistically significant.

For female adult speakers, the performance differences between the baseline and the VTLN (from 10.91% to 10.41% PER), between the MFCC features vector augmented with the posterior probabilities of the warping factors and the baseline (from 10.91% to 10.89% PER) and between VTLN and DNN-adaptation (from 10.41% to 9.65% PER) are not statistically significant.

For male adult speakers, the system working with the MFCC features vector augmented with the posterior probabilities of the warping factors improves the baseline by 3% relative (from 8.62% to 8.34% PER with $p < .01$). The VTLN normalisation approach improve the baseline performance by 9% relative (from 8.62% to 7.91% PER with $p < .001$) and the performance difference between VTLN and DNN-adaptation (from 7.91% to 7.61% PER) is not statistically significant.

	Evaluation Set							
	ChildIt		APASCI (f)		APASCI (m)		ChildIt + APASCI	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
Baseline	13.98%	15.56%	10.12%	10.91%	10.07%	8.62%	12.47%	14.29%
DNN-adaptation	11.68%	12.43%	8.30 %	9.65%	9.33%	7.61%	10.39%	11.59%
VTLN-normalisation	11.94%	12.80%	9.06%	10.41%	9.76%	7.91%	10.81%	12.00%
VLTN + MFCC	12.67 %	14.10%	9.02 %	10.89%	9.75%	8.34%	11.21%	13.12%

Table 1: Phone error rate achieved with VTLN approaches to DNN-HMM.

5.2. System integration and complexity

When considering the integration to a complete system, the DNN adaptation approach requires to train three age/gender group-specific DNN. At runtime, two modalities can be adopted: a) model selection which requires the use of a pre-trained a classifier to select the proper DNN-HMM system for each utterance to decode, b) multiple decodings with the three age/gender group-specific DNN-HMM systems and keeping the output with the highest likelihood. In the DNN adaptation approach, if a target group of speakers is changed or added, there is the need to train a new DNN corresponding to the new target group of speakers. Approaches relying on VTLN are more general in this sense. At runtime, normalising the MFCC with VTLN requires a two-pass decoding system which is unsuited for online applications. The approach based on MFCC features vector augmented with the posterior probabilities of the warping factors relies on only one DNN-HMM system and one small DNN to obtain the posterior probabilities. It can operate in one-pass and it is the simplest system presented here. Besides, this latter approach relies only on the DNN and it would allow to perform a joint optimisation of the whole system at once (DNN-warp and DNN-HMM) in a similar way as in [30]. Therefore, each of the systems compared here can fit different scenarios: DNN-adaptation when computational resources for decoding is not limited, VTLN applied to the MFCC when considering off-line decoding, and MFCC augmented with the posterior probabilities of the warping factors when a small, flexible system is needed.

6. CONCLUSIONS

In this paper we have investigated the use of two VTLN approaches for DNN-HMM in a phone recognition task when targeting a heterogeneous speaker population consisting of children, adult males and adult females. Performance obtained with these approaches were compared with earlier work on DNN-adaptation. When only limited training data is available, normalising the MFCC through VTLN for a DNN-HMM system can help to improve the performance by up to 20% relative compared to the baseline. The system operating on VTLN normalised MFCC then performs almost as well as the DNN-adaptation with matched training and testing conditions.

An alternative approach has been presented in which MFCC derived acoustic features are combined with the posterior probabilities of the VTLN warping factors to obtain an augmented set of features as input to a DNN. This approach has been shown to perform slightly worse than conventional VTLN applied to DNN-HMM but it still allows to improve PER performance by up to 10% relative compared to the baseline. Besides, this approach is the simplest of the three approaches compared here and the fact that it relies only on DNN makes it promising for future developments such as joint optimisation of the DNN-warp and the DNN-HMM.

7. REFERENCES

[1] W. T. Fitch and J. Giedd, "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," *Journal of Acoust. Soc. Amer.*, vol. 106, no. 3, pp. 1511–1522, Sept. 1999.

[2] J. E. Huber, E. T. Stathopoulos, G. M. Curione, T. A. Ash, and K. Johnson, "Formants of children women and men: The effect of vocal intensity variation," *Journal of Acoust. Soc. Amer.*, vol. 106, no. 3, pp. 1532–1542, Sept. 1999.

[3] S. Lee, A. Potamianos, and S. Narayanan, "Acoustic of children's speech: Developmental changes of temporal and spectral parameters," *Journal of Acoust. Soc. Amer.*, vol. 105, no. 3, pp. 1455–1468, March 1999.

[4] L. Lee and R. C. Rose, "Speaker Normalization Using Efficient Frequency Warping Procedure," in *Proc. of IEEE ICASSP*, Atlanta, GA, May 1996, pp. 353–356.

[5] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin, "Speaker Normalisation on Conversational Telephone Speech," in *Proc. of IEEE ICASSP*, Atlanta, GA, May 1996, pp. I–339–341.

[6] E. Eide and H. Gish, "A Parametric Approach to Vocal Tract Lenght Normalization," in *Proc. of IEEE ICASSP*, Atlanta, GA, May 1996, pp. 346–349.

[7] A. Hagen, B. Pellom, and R. Cole, "Children's Speech Recognition with Application to Interactive Books and Tutors," in *Proc. of IEEE ASRU Workshop*, St. Thomas Irsee, US Virgin Islands, Dec. 2003.

[8] D. Giuliani and M. Gerosa, "Investigating Recognition of Children Speech," in *Proc. of IEEE ICASSP*, vol. 2, Hong Kong, Apr. 2003, pp. 137–140.

[9] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*. Springer, 1994, vol. 247.

[10] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[11] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, Jan 2012.

[12] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, Jan 2012.

[13] A. Metallinou and J. Cheng, "Using Deep Neural Networks to Improve Proficiency Assessment for Children English Language Learners," in *Proc. of INTERSPEECH*, 2014, pp. 1468–1472.

[14] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. of IEEE ASRU Workshop*, December 2011.

[15] R. Serizel and D. Giuliani, "Deep neural network adaptation for children's and adults' speech recognition," in *Proc. of the First Italian Computational Linguistics Conference*, Pisa, IT, 2014.

[16] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proc. of IEEE SLT Workshop*, Dec 2012, pp. 246–251.

[17] V.-B. Le, L. Lamel, and J. Gauvain, "Multi-style ML features for BN transcription," in *Proc. of IEEE ICASSP*, March 2010, pp. 4866–4869.

[18] S. Thomas, M. Seltzer, K. Church, and H. Hermansky, "Deep neural network features and semi-supervised training for low resource speech recognition," in *Proc. of IEEE ICASSP*, May 2013, pp. 6704–6708.

[19] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. of IEEE ICASSP*, 2013, pp. 7947–7951.

[20] O. Abdel-Hamid and H. Jiang, "Rapid and effective speaker adaptation of convolutional neural network based models for speech recognition." in *Proc. of INTERSPEECH*, 2013, pp. 1248–1252.

[21] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proc. of IEEE ICASSP*, 2013.

[22] A. Senior and I. Lopez-Moreno, "Improving DNN speaker independence with I-vector inputs," in *Proc. of IEEE ICASSP*, 2014.

[23] L. Welling, S. Kanthak, and H. Ney, "Improved Methods for Vocal Tract Normalization," in *Proc. of IEEE ICASSP*, vol. 2, Phoenix, AZ, April 1999, pp. 761–764.

[24] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of childrens speech," *Speech Communication*, vol. 49, no. 1011, pp. 847 – 860, 2007.

[25] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo, "Speaker Independent Continuous Speech Recognition Using an Acoustic-Phonetic Italian Corpus," in *Proc. of ICSLP*, Yokohama, Japan, Sept. 1994, pp. 1391–1394.

[26] K. Veselỳ, L. Burget, and F. Grézl, "Parallel training of neural networks for speech recognition," in *Text, Speech and Dialogue*. Springer, 2010, pp. 439–446.

[27] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[28] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pretraining help deep learning?" *The Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.

[29] L. Gillick and S. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms," in *Proc. of IEEE ICASSP*, Glasgow, Scotland, May 1989, pp. I–532–535.

[30] A. Narayanan and D. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Proc. of IEEE ICASSP*, 2014, pp. 2523–2527.