Deep-neural network approaches for speech recognition with heterogeneous groups of speakers including children

ROMAIN SERIZEL
Institut Mines-Télécom/Télécom ParisTech, CNRS-LTCI
37, rue Dareau, Paris, France 75014
and
HLT research unit, Fondazione Bruno Kessler (FBK)
Via Sommarive 18, Trento, Italy 38121

and DIEGO GIULIANI
HLT research unit, Fondazione Bruno Kessler (FBK)
Via Sommarive 18, Trento, Italy 38121

(*Received 23 February 2016*)

Abstract

This paper introduces deep neural network (DNN) - hidden Markov model (HMM) based methods to tackle speech recognition in heterogeneous groups of speakers including children. We target three speaker groups consisting of children, adult males and adult females. Two different kind of approaches are introduced here: approaches based on DNN adaptation and approaches relying on vocal-tract length normalisation (VTLN).

First, the recent approach that consists in adapting a general DNN to domain/language specific data is extended to target age/gender groups in the context of DNN-HMM. Then, VTLN is investigated by training a DNN-HMM system by using either mel frequency cepstral coefficients (MFCC) normalised with standard VTLN or MFCC derived acoustic features combined with the posterior probabilities of the VTLN warping factors. In this later, novel, approach the posterior probabilities of the warping factors are obtained with a separate DNN and the decoding can be operated in a single pass when the VTLN approach requires two decoding passes. Finally, the different approaches presented here are combined to take advantage of their complementarity. The combination of several approaches is shown to improve the baseline phone error rate performance by 30% to 35% relative and the baseline word error rate performance by about 10% relative.

1 Introduction

Speaker-related acoustic variability is a 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 DNN is used 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 (Fitch and Giedd, 1999) 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 (Huber, Stathopoulos, Curione, Ash, and Johnson, 1999; Lee, Potamianos, and Narayanan, 1999). 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 substantial longer vocal tract and lower formant frequencies than females. Consequently, voices of children tend to be more similar to the voices of women than to those of men.

When an automatic speech recognition (ASR) system trained on adults' speech is employed to recognise children's speech, performance decreases drastically, especially for younger children (Wilpon and Jacobsen, 1996; Claes, Dologlou, ten Bosch, and Compernolle, 1998; Das, Nix, and Picheny, 1998; Li and Russell, 2001; Giuliani and Gerosa, 2003; Potamianos and Narayanan, 2003; Gerosa, Giuliani, and Brugnara, 2007; Gerosa, Giuliani, Narayanan, and Potamianos, 2009b). A number of attempts have been reported in the literature to compensate for this effect. Most of them try to compensate for spectral differences caused by differences in vocal tract length and shape by warping the frequency axis of the speech power spectrum of each test speaker or transforming acoustic models (Claes et al., 1998; Das et al., 1998; Potamianos and Narayanan, 2003). However, to ensure good recognition performance, age-specific acoustic models trained on speech collected from children of the target age, or group of ages, is usually employed (Wilpon and Jacobsen, 1996; Hagen, Pellom, and Cole, 2003; Nisimura, Lee, Saruwatari, and Shikano, 2004; Gerosa et al., 2007). Typically much less training data are available for children than for adults. The use of adults' speech for reinforcing the training data in the case of a lack of children's speech was investigated in the past (Wilpon and Jacobsen, 1996; Steidl, Stemmer, Hacker, Nöth, and Niemann, 2003). However, in order to achieve a recognition performance improvement when training with a mixture of children's and adults' speech, speaker normalisation and speaker adaptive training techniques are usually needed (Gerosa, Giuliani, and Brugnara, 2009a).

How to cope with acoustic variability induced by gender differences has been studied for adult speakers in a number of papers. Assuming that there is enough training data, one approach consists in the use of gender-dependent models that are either directly used in the recognition process itself (Yochai and Morgan, 1992; Woodland, Odell, Valtchev, and Young, 1994) or used as a better seed for speaker adaptation (Lee and Gauvain, 1993). Alternatively, when training on speakers of both genders, speaker normalisation and adaptation techniques are commonly employed to compensate for acoustic inter-speaker variability (Lee and Rose, 1996; Gales, 1998).

Since the surfacing of efficient pre-training algorithms during the past years (Hinton, Osindero, and Teh, 2006; Bengio, Lamblin, Popovici, and Larochelle, 2007;

Erhan, Bengio, Courville, Manzagol, Vincent, and Bengio, 2010; Seide, Li, Chen, and Yu, 2011), DNN has proven to be an effective alternative to Gaussian mixture model (GMM) in HMM-GMM based ASR (Bourlard and Morgan, 1994; Hinton, Deng, Yu, Dahl, Mohamed, Jaitly, Senior, Vanhoucke, Nguyen, Sainath, and Kingsbury, 2012) and really good performance has been obtained with hybrid DNN-HMM systems (Dahl, Yu, Deng, and Acero, 2012; Mohamed, Dahl, and Hinton, 2012).

Capitalising on their good classification and generalisation capabilities DNNs have been used widely in multi-domain and multi-languages tasks (Sivadas and Hermansky, 2004; Stolcke, Grezl, Hwang, Lei, Morgan, and Vergyri, 2006). The main idea is usually to first exploit a task independent (multi-lingual/multi-domain) corpus and then to use a task specific corpus. These different corpora can be used to design new DNN architectures with application to task specific ASR (Pinto, Magimai-Doss, and Bourlard, 2009) or task independent ASR (Bell, Swietojanski, and Renals, 2013). Another approach consists in using the different corpora at different stages of the DNN training. The task independent corpus is used only for the pre-training (Swietojanski, Ghoshal, and Renals, 2012) or for a general first training (Le, Lamel, and Gauvain, 2010; Thomas, Seltzer, Church, and Hermansky, 2013) and the task specific corpus is used for the final training/adaptation of the DNN. In under-resourced scenarios, approaches based on DNN (Imseng, Motlicek, Garner, and Bourlard, 2013) have then shown to outperform approaches based on subspace GMM (Burget, Schwarz, Agarwal, Akyazi, Feng, Ghoshal, Glembek, Goel, Karafiat, Povey, Rastrow, Rose, and Thomas, 2010).

However, to our best knowledge, apart from the very recent work on the subject in Metallinou and Cheng (2014) DNN is scarcely used in the context of children's speech recognition. In Wöllmer, Schuller, Batliner, Steidl, and Seppi (2011) a bidirectional long short-term memory network is used for keyword detection but we have not found any mention of the application of the hybrid DNN-HMM to children's speech recognition.

Three target groups of speakers are considered in this work, that is children, adult males and adult females. There is only a limited amount of labelled data for such groups. We investigated two approaches for ASR in under-resourced conditions with an heterogeneous population of speakers.

The first approach investigated in this paper extends the idea introduced in Yochai and Morgan (1992) to the DNN context. The DNN trained on speech data from all the three groups of speakers is adapted to the age/gender group specific corpora. First it is shown that training a DNN only from a group specific corpus is not effective when only limited labelled data is available. Then the method proposed in Thomas et al. (2013) is adapted to the age/gender specific problem and used in a DNN-HMM architecture instead of a tandem architecture.

The second approach introduced in this paper relies on VTLN. In Seide et al. (2011) 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 un-normalised acoustic features. It was argued that, when a sufficient amount of training data is available, DNNs 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 to reach strong generalisation capabilities (Serizel and Giuliani, 2014a). In such case, techniques like DNN adaptation (Le et al., 2010; Swietojanski et al., 2012; Thomas et al., 2013), speaker adaptation (Abdel-Hamid and Jiang, 2013b; Liao, 2013) or VTLN (Eide and Gish, 1996; Lee and Rose, 1996; Wegmann, McAllaster, Orloff, and Peskin, 1996) can help to improve the performance. Here we consider first the application of a conventional VTLN technique to normalise MFCC vectors as input features to a DNN-HMM system.

Recent works have shown that augmenting the inputs of a DNN with, e.g. an estimate of the background noise (Seltzer, Yu, and Wang, 2013) or utterance i-vector (Senior and Lopez-Moreno, 2014), 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.

This paper extends previous work by the authors on DNN adaptation (Serizel and Giuliani, 2014a) and VTLN approaches for DNN-HMM based ASR (Serizel and Giuliani, 2014b). An approach to optimise jointly the DNN that extracts the posterior probabilities of the warping factors and the DNN-HMM is proposed here, combination of the different approaches is considered and performance of the different systems are evaluated not only on phone recognition but also on word recognition.

This paper is a proof of concept and its scope is limited to the investigation of a simple acoustic model adaptation approach and several VTLN related approaches. To cope with inter-speaker acoustic variability induced by age and gender, state-of-the-art approaches based on speaker identity models such as I-vectors (Dehak, Kenny, Dehak, Dumouchel, and Ouellet, 2011; Saon, Soltau, Nahamoo, and Picheny, 2013; Senior and Lopez-Moreno, 2014), speaker codes (Abdel-Hamid and Jiang, 2013a), linear input networks and linear output networks (Li and Sim, 2010) could be considered although they are beyond the scope of this paper.

The rest of the paper is organised as follows, Section 2 briefly introduces DNNs for acoustic modelling in ASR and presents the approach based on DNN adaptation. Approaches based on VTLN are presented in Section 3. The experimental set-up is described in Section 4 and experiments results are presented in Section 5. Finally, conclusions of the paper are drawn in Section 6.

2 DNN adaptation

A DNN is a feed-forward neural network where the neurons are arranged in fully connected layers. The input layer processes the feature vectors (augmented with context) and the output layer provides (in the case of ASR) the posterior probability of the (sub)phonetic units. The layers between the input layer and the output layer are called hidden layers. DNNs are called deep because they are composed of many layers. Even though shallow neural network architectures (i.e., with few hidden layers) are supposed to be able to model any function, they may require a huge number of parameters to do so. The organisation of the neurons in a deep architecture allows to use parameters more efficiently and to model the same function as a shallow architectures with less parameters (Bengio, Courville, and Vincent, 2013). Deep architectures also allow to extract high level features that are more invariant (and therefore more robust) than low level features (Hinton et al., 2012). They also allow to close the semantic gap between the features and the (sub)phonetic units.

The DNN used in this papers have sigmoid activation functions in the hidden layers:

$$h = \mathbf{w}.\mathbf{y} + b$$

$$\sigma(h) = \frac{1}{1 + e^{-h}}$$

with $\mathbf{y}$ the vector of input to the layer, $\mathbf{w}$ and b the weights and the bias of a given neuron in the layer, respectively.

The target of the DNN presented here is to estimate posteriors probabilities. Therefore, it is chosen to use softmax activation in the output layer, as the outputs then sum up to one:

$$softmax(h_j) = \frac{e^{h_j}}{\sum\limits_i e^{h_i}}$$

with i running over the neurons in the output layer.

The state posterior probabilities are then normalised by the state prior probabilities to obtain the state emission likelihood used by the HMM. Following Bayes' rule:

$$p(X|S) \propto \frac{p(S|X)}{p(S)}$$

where X is the acoustic observation and S the HMM state.

2.1 Pre-training/training procedure

Training a DNN is a difficult task mainly because the optimisation criterion involved is non convex. Training a randomly initialised DNN with back-propagation would converge to one of the many local minima involved in the optimisation problem, sometimes leading to poor performance (Erhan et al., 2010). In recent works this limitation has been partly overcome by training on a huge amount of data (1700 hours in Senior and Lopez-Moreno (2014)). However, this solution does not apply when tackling ASR for under-resourced groups of population where the amount of training data is limited by definition. In such cases, pre-training is a mandatory step to efficiently train a DNN. The aim of pre-training is to initialise the DNN weights to a better starting point than randomly initialised DNN and avoid the back-propagation training to be stuck in a poor local minima. Here generative training based on Restricted Boltzmann Machines (RBM) (Hinton et al., 2006; Erhan et al., 2010) is chosen. Once the DNN weights have been initialised with stacked RBM, the DNN is trained to convergence with back-propagation. More details about training and network parameters are presented in Sections 4.2.2 and 4.3.2.

2.2 Age/gender independent training

The general training procedure described above can be applied, by using all training data available, in an attempt to achieve a system with strong generalisation capabilities. Estimating the DNN parameters on speech from all groups of speakers, that is children, adult males and adult females, 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.

2.3 Age/gender adaptation

ASR systems provide their best recognition performances when the operating (or testing) conditions match the training conditions. To be effective, the general training procedure described above requires that a sufficient amount of labelled data is available. Therefore, when considering training for under-resourced population groups (such as children or males/females in particular domains of applications) it might be more effective to train first a DNN on all data available and then to adapt this DNN to a specific group of speakers. A similar approach has been proposed in Thomas et al. (2013) for the case of multilingual training. In this paper the language does not change and the targets of the DNN remain the same when going from age/gender independent training to group specific adaptation. The DNN trained on speech data from all groups of speakers can then be used directly as initialisation to the adaptation procedure where the DNN is trained to convergence with back-propagation only on group specific speech corpora.

This adaptation approach, however, suffers from a lack of flexibility: a new DNN would have to be adapted to each new group of speakers.

3 VTLN approaches

In this section, we propose to define a more general framework inspired by VTLN approaches to ASR to tackle the problem of inter-speaker acoustic variability due to vocal tract length (and shape) variations among speakers. Two different approaches are considered here. The first one is based on the conventional VTLN approach (Eide and Gish, 1996; Lee and Rose, 1996; Wegmann et al., 1996). The resulting VTLN normalised acoustic features are used as input to the DNN both during training and testing (Seide et al., 2011). 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 un-normalised acoustic feature vector, extended with context, to form an augmented acoustic feature 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 (Eide and Gish, 1996; Lee and Rose, 1996; Wegmann et al., 1996), typical issues are

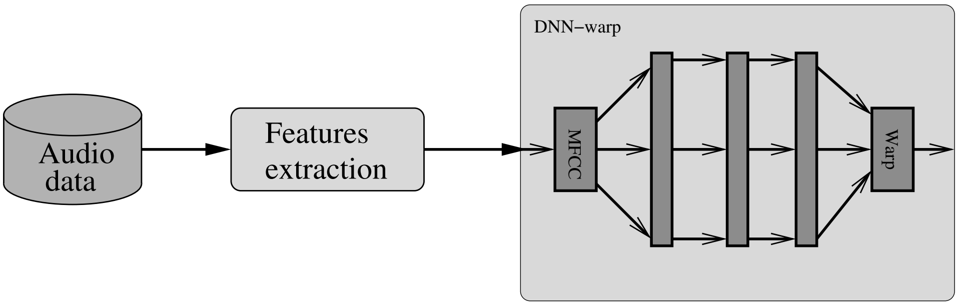

Fig. 1. Training of the DNN-warp.

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 its transcription and a current set of HMM-based acoustic models (Lee and Rose, 1996). 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 (Lee and Rose, 1996). In this work we adopted the latter approach considering a discrete set of VTLN factors. Details on the VTLN implementation are provided in Section 4.5.

Similarly to the method proposed in Seide et al. (2011), 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 feature 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 to noise and speaker independence of the DNN (Seltzer et al., 2013; Senior and Lopez-Moreno, 2014).

The VTLN procedure is first applied to generate a warping factor for each utterance in the training set. Each acoustic feature vector in the utterance is labelled with the utterance warping factor. Then, training acoustic feature vectors and corresponding warping factors are used to train a DNN classifier. Each class of the DNN correspond to one of the discrete VTLN factors and the dimension of the DNN output corresponds to the number of discrete VTLN factors. The DNN learns to infer the VTLN warping factor from the acoustic feature vector (Figure 1) or more precisely the posterior probability of each VTLN factor knowing the input acoustic feature vector. This DNN will be referred to as DNN-warp.

During training and testing of the DNN-HMM system, for each speech frame the warping factors posterior probabilities are estimated with the DNN-warp. These es-

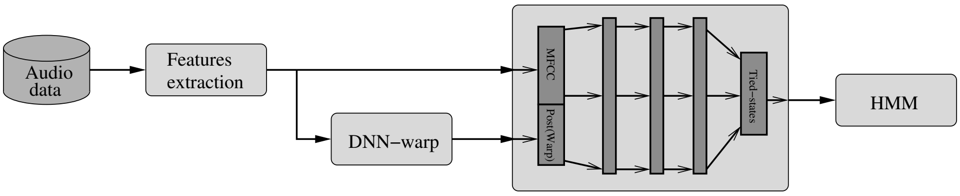

Fig. 2. Training of the warping factor aware DNN-HMM.

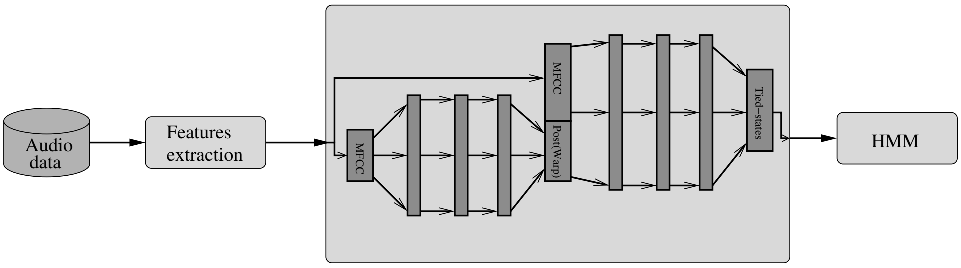

Fig. 3. Joint optimisation of the DNN-warp and the DNN-HMM.

timated posterior probabilities are appended to the un-normalised acoustic feature vectors, extended with context, to form an augmented acoustic feature vectors. Mean and variance normalisation is then applied to the extended feature vector which is used as input to the DNN-HMM (Figure 2).

This approach has the advantage to reduce considerably the complexity during decoding compared to the approach making use of conventional VTLN normalised acoustic features that requires a preliminary decoding pass to obtain a transcript of acoustic data to be used for estimating the warping factor (Lee and Rose, 1996; Welling, Kanthak, and Ney, 1999). It also allows for flexible estimation of the warping factors: they could either be updated on a frame to frame basis or averaged at utterance level (see also Section 5).

3.3 Joint optimisation

The ultimate goal here is not to estimate the VTLN warping factors but to perform robust speech recognition on heterogeneous corpora. To this end, the DNN-warp and the DNN-HMM can be optimised jointly (Figure 3). The procedure is the following one: 1) first the DNN-warp is trained alone (Figure 1); 2) the posteriors of the warping factors on the training set are obtained with the DNN-warp; 3) these posteriors of the warping factors are used as input to the DNN-HMM together with the acoustic features to produce an extended feature vector; 4) the DNN-HMM is trained (Figure 2); 5) the DNN-warp and the DNN-HMM are concatenated to obtained a deeper network that is fine-tuned with back-propagation on the training set (Figure 3). Details about joint optimisation are presented in Section 4.6

	Speech Corpus				
	ChildIt	APASCI(f)	APASCI(m)	IBN(f)	IBN(m)
Train	7h:15m	2h:40m	2h:40m	23h:00m	25h:00m
Test	2h:20m	0h:20m	0h:20m	1h:00m	1h:00m

Table 1. *Data repartition in the speech corpora. (f) and (m) denote speech from female and male speakers, respectively.*

	Grade						
	2	3	4	5	6	7	8
N. Speakers	24	24	23	24	28	26	22

Table 2. *Distribution of speakers in the ChildIt corpus per grade. Children in grade 2 are approximatively 7 years old while children in grade 8 are approximatively 13 years old.*

4 Experimental set-up

4.1 Speech corpora

For this study we relied on three Italian speech corpora: the ChildIt corpus consisting of children's speech, the APASCI corpus and the IBN corpus consisting of adults' speech. All corpora were used for evaluation purposes, while ChildIt and APASCI provide a similar amount of training data for children and adults, respectively, the IBN corpus contains approximately 5 times as much training data as ChildIt or APASCI (Table 1).

4.1.1 ChildIt

The ChildIt corpus (Giuliani and Gerosa, 2003; Gerosa et al., 2007) consists of Italian read sentences collected from 171 children (86 male and 85 female) aged between 7 and 13, with a mean age of 10 years. Recordings took place at school, usually in the computer room or in the library. Each child was asked to read a set of sentences prepared according to her/his grade. Figure 2 reports the distribution of children per grade.

The overall duration of audio recordings in the corpus is 10h:24m. For all recordings in the corpus a word-level transcription is available.

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, for a total durations of 0h:49m; a test set consisting of data from 42 speakers balanced with respect to age and gender for a total duration of of 2h:20m.

4.1.2 APASCI

The APASCI speech corpus (Angelini, Brugnara, Falavigna, Giuliani, Gretter, and Omologo, 1994) is a task-independent, high quality, acoustic-phonetic Italian corpus. APASCI consists of read speech collected from 194 adult speakers for a total durations of 7h:05m. For all recordings in the corpus a word-level transcription is available. The corpus is 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; a test set consisting of data from 30 speakers balanced per gender, for a total duration of 0h:40m.

4.1.3 IBN Corpus

The IBN corpus is composed of speech from several radio and television Italian news programs (Gerosa et al., 2009a). It consists of adult speech only, with word-level transcriptions. The IBN corpus was partitioned into a training set, consisting of 52h:00m of speech, and a test set formed by 2h:00m of speech. During the experiments presented here 2h:00m of male speech and 2h:00m of female speech are extracted from the training set to be used as development set during the DNN training. The resulting training set is then partitioned into 25h:00m of male speech and 23h:00m of female speech.

4.2 Phone recognition systems

The approaches proposed in this paper have been first tested on small corpora (ChildIt + APASCI) for phone recognition to explore as many set-ups as possible in a limited amount of time. The reference phone transcription of an utterance was derived from the corresponding word transcription by performing Viterbi decoding on a pronunciation network. This pronunciation network was built by concatenation of the phonetic transcriptions of the words in the word transcription. In doing this alternative word pronunciations were taken into account and an optional insertion of the silence model between words was allowed.

4.2.1 GMM-HMM

The acoustic features are 13 mel frequency cepstral coefficients (MFCC), including the zero order coefficient, computed on 20ms frames with 10ms overlap. First, second and third order time derivatives are computed after cepstral mean subtraction performed utterance by utterance. These features are arranged into a 52-dimensional vector that is projected into a 39-dimensional feature space by applying a linear transformation estimated through Heteroscedastic Linear Discriminant Analysis (HLDA) (Kumar and Andreou, 1998).

Acoustic models are 3039 tied-state triphone HMMs based on a set of 48 phonetic units derived from the SAMPA Italian alphabet. Each tied-state is modelled with a mixture of 8 Gaussian densities having a diagonal covariance matrix. In addition, "silence" is modelled with a Gaussian mixture model having 32 Gaussian densities.

4.2.2 DNN-HMM

The DNN uses again 13 MFCC, including the zero order coefficient, computed on 20ms frames with 10ms overlap. The context spans on a 31 frames window. For each frequency band, the 31 coefficients context is separately scaled with a Hamming window and projected to a 16 dimensional vector using DCT. The 13 resulting vectors are concatenated to obtain a 208 dimensional feature vector which is normalised to have zero-mean and unit variance before being used as input to the DNN. The targets of the DNN are the 3039 tied-states obtained from the GMM-HMM training 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 (Veselỳ, Burget, and Grézl, 2010). The DNN weights are initialised randomly and pre-trained with RBM. 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. The values of the hyper-parameters (network topology and learning parameters) are standard values, in the range of the values commonly used for these parameters in the literature. Considering the relatively small size of the corpora, the number of hidden layers is set to 4. Increasing the number of layers with the amount of data available has been observed to provide no significant performance improvement. Besides, training a system with more than 6 hidden layers will result in lower performance than with 4 hidden layers.

The DNN can be trained either on all speech data available (ChildIt + APASCI) or on group specific corpora (ChildIt, adult female speech in APASCI, adult male speech in APASCI).

4.2.3 Language model

A simple finite state network having just one state and a looped transition for each phone unit was employed. In this network uniform transition probabilities are associated to looped transitions. In computing recognition performance, in terms of PER, no distinction was made between single consonants and their geminate counterparts. In this way, the set of phonetic labels was reduced from 48 to 28 phone labels.

4.3 Word recognition systems

The approaches that performed best in phone recognition on the small corpora are validated in word recognition on a more realistic set-up (ChildIt+ IBN) including a corpus of adult speech (IBN) that is larger than the corpus of children speech (ChildIt).

4.3.1 GMM-HMM

The GMM-HMM are similar to those used for phone recognition except that they use more Gaussian densities to benefit from the extensive training data. Acoustic models are 5021 tied-state triphone HMM based on a set of 48 phonetic units derived from the SAMPA Italian alphabet. Each tied-state is modelled with a mixture of 32 Gaussian densities having a diagonal covariance matrix. In addition, "silence" is modelled with a Gaussian mixture model having 32 Gaussian densities.

4.3.2 DNN-HMM

The DNN are similar to those used for phone recognition except that they are trained on a different set of targets. The targets of the DNN are the 5021 tied-states obtained from the word recognition GMM-HMM training on the mixture of adults' and children's speech (ChildIt + IBN). 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 5021.

4.3.3 Language model

For word recognition, a 5-gram language model was trained on texts from the Italian news domain consisting of about 1.6G words. Part of the textual data, consisting in about 1.0G words, were acquired via web crawling of news domains. The recognition dictionary consists of the most frequent 250K words.

4.4 Age/gender adapted DNN for DNN-HMM

One option is to adapt an already trained general DNN to group specific corpora. The data architecture is the same as described above. The initial DNN weights are the weights obtained with a pre-training/training procedure applied on all training data available (ChildIt+APASCI, respectively ChildIt + IBN). The DNN is then trained with back propagation on a group specific corpus (ChildIt, adult female speech in APASCI and adult male speech in APASCI, respectively IBN). The training parameters are the same as during the general training (4.2.2 and 4.3.2, respectively) and the learning rate follows the same rule as above. The mini-batch size is 512 and a first-order momentum of 0.5 is applied.

4.5 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, as proposed in (Welling et al., 1999), and trained on un-warped children's and adults' speech (Gerosa et al., 2007, 2009a).

The DNN-warp inputs are the MFCC with a 61 frames context window, DCT projected to a 208 dimensional feature vector (the procedure is similar as in 4.2.2). 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-dimensional DCT projected acoustic feature vector to produce a 233-dimensional feature vector that is mean-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 for phone recognition and 233 x 1500 x 1500 x 1500 x 1500 x 5021 for word recognition.

4.6 Joint optimisation

The DNN-warp and DNN-HMM can be fine-tuned jointly with back-propagation. In such case, the starting learning rate is set to 0.0002 in the first 4 hidden layers (corresponding to the DNN-warp) and to 0.0001 in the last 4 hidden layers (corresponding to the DNN-HMM). The learning rate is chosen empirically as the highest value for which both training accuracy and cross-validation accuracy improve. Setting a different learning rate in the first 4 hidden layers and the last 4 hidden layers is done in an attempt to overcome the vanishing gradient effect in the 8 layers DNN obtained from the concatenation of the DNN-warp and the DNN-HMM. The learning rates are then adapted following the same schedule as described above. The joint optimisation is done with a modified version of the TNet software package (Veselỳ et al., 2010).

5 Experimental Results

Two sets of experiments are presented here. First the systems are tested extensively in terms of PER on small corpora (ChildIt + APASCI), then the best performing systems are tested in terms of WER performance on a more realistic set-up including a larger adult speech corpus (IBN).

5.1 Phone recognition

The experiments presented here are designed to verify the validity of the following statements:

- The age/gender group specific training of the DNN does not necessarily lead to improved performance, specially when a small amount of data is available
- The age/gender group adaptation of a general DNN can help to design group specific systems, even when only a small amount of data is available
- VTLN can be beneficial to the DNN-HMM framework when targeting a heterogeneous speaker population with limited training data
- Developing an "all-DNN" approach to VTLN for a 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
- Optimising the DNN-warp and the DNN-HMM jointly can help to improve the performance in certain cases
- The different approaches introduced in this paper can be complementary.

During the experiments the language model weight is tuned on the development set and used to decode the test set. Results were obtained with a phone loop language model and the PER was computed based on 28 phone labels. Variations in recognition performance were validated using the matched-pair sentence test (Gillick and Cox, 1989) 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.

5.1.1 Age/gender specific training for DNN-HMM

In this experiment, DNNs are trained on group specific corpora (children's speech in ChildIt, adult female speech in APASCI and adult male speech in APASCI) and performance is compared with the DNN-HMM baseline introduced above where the DNN is trained on speech from all speaker groups. Recognition results are reported in Table 3, which includes results achieved with the DNN-HMM baseline in the row *Baseline*. In ChildIt there is about 7h of training data which is apparently sufficient to train an effective DNN and we can observe an improvement of 22% PER relative compared to the baseline performance (from 15.56% to 12.76% with $p < .001$). However, in adult data there is only about 2h:40m of data for each gender. This is apparently not sufficient to train a DNN. In fact, the DNN-HMM system based on a DNN that is trained on gender specific data consistently degrades the PER. The degradation compared to the baseline performance is 14% PER relative on female speakers in APASCI (from 10.91% to 12.75% with $p < .001$) and 12% PER relative on male speakers in APASCI (from 8.62% to 9.83% with $p < .001$).

5.1.2 Age/gender adapted DNN-HMM

In this experiment the DNN trained on all available corpora is adapted to each group specific corpus and recognition performance is compared with that obtained by the

	Evaluation Set		
Training Set	ChildIt	APASCI(f)	APASCI(m)
Baseline	15·56%	**10·91%**	**8·62%**
ChildIt	**12·76%**	29·59%	46·16%
APASCI(f)	34·23%	12·75%	31·21%
APASCI(m)	56·11%	30·81%	9·83%

Table 3. *Phone error rate achieved with the DNN-HMM trained age/gender groups specific data.*

DNN-HMM baseline (where the DNN is trained on all available corpora). PER performance is presented in Table 4 which also reports the results achieved by the DNN-HMM baseline (in row *Baseline*). The group adapted DNN-HMM consistently improve the PER compared to the DNN-HMM baseline. On children's speech the PER improvement compared to the baseline is 25% PER relative (from 15.56% to 12.43% with $p < .001$). On adult female speakers in APASCI the age/gender adaptation improves the baseline performance by about 13% PER relative (from 10.91% to 9.65% with $p < .001$). On adult male speakers the age/gender adaptation improves the baseline performance by 13% (from 8.62% to 7.61% with $p < .05$).

From the results in Table 4 it is also possible to note that the DNN-HMM system adapted to children's voices performs much better for adult female speakers than for adult male speakers. Similarly, the DNN-HMM system adapted to female voices perform better on children' speech than the system adapted to male voices. These results are consistent with results in Table 3 and confirm that characteristics of children's voice is much more similar to those of adult female voices than those of adult male voices.

In the *Model selection (oracle)* approach, we assumed that a perfect age/gender classifier exist which allows us to know in which target group of speaker an incoming speech segment belongs. The recognition is then performed using the corresponding adapted model. On the evaluation set including all the target groups of speakers (ChildIt + APASCI) the use of matched acoustic models improves the baseline by 23% PER relative (from 14.32% to 11.59% with $p < .05$).

For comparison purposes, last row of Table 4 (*Model selection*) reports results obtained with an automatic approach for acoustic model selection. In this case each utterance is decoded three times by using each individual group adapted acoustic model and, as final recognition result, the recognition hypothesis resulting in the highest likelihood is retained. Comparing recognition results in the last two rows of

	Evaluation Set			
Adaptation Set	ChildIt	APASCI(f)	APASCI(m)	ChildIt + APASCI
Baseline	15·56%	10·91%	8·62%	14·32%
ChildIt	**12·43%**	16·93%	24·96%	N/A
APASCI(f)	21·91%	**9·65%**	17·01%	N/A
APASCI(m)	32·33%	16·99%	**7·61%**	N/A
Model selection (oracle)	12·43%	9·65%	7·61%	11·59%
Model selection	12·97%	10·98%	8·49%	12·26%

Table 4. *Phone error rate achieved with the DNN-HMM trained on a mixture of adult and children's speech and adapted to specific age/gender groups.*

Table 4 it is possible to note that the automatic model selection approach results in an overall decrease of performance: from 11.59% to 12.26% PER. This decrease of performance is consistent across the three groups of speakers. It would probably be possible to obtain better model selection for example by training a DNN to perform the selection but this is out of the scope of this paper. Therefore, in the rest of the paper, *Model selection* approach is assumed to be the *Model selection (oracle)* approach and recognition experiments are always conducted with matching adapted acoustic models.

5.1.3 VTLN based approaches

Table 5 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*), the MFCC feature vector augmented with the the warping factors obtained in a standard way (row *Warp + MFCC*), the MFCC features augmented with the posterior probabilities of the warping factors (row *Warp-post + MFCC*), the MFCC features augmented with the posterior probabilities of the warping factors averaged at utterance level (row *Warp-post (utt) + MFCC*) and the joint optimisation of the DNN-warp and the DNN-HMM (row *Warp-post + MFCC (joint)*).

To compute the vectors *Warp-post (utt) + MFCC* the posterior probability of each warping factor is averaged over utterances to obtain a vector of averaged posterior probabilities. This experiment allows to study independently the effects of having a soft or hard decision on the warping factor selection and the effects of

	Evaluation Set			
	ChildIt	APASCI(f)	APASCI(m)	ChildIt + APASCI
Baseline	15·56%	10·91%	8·62%	14·32%
VTLN-normalisation	12·80%	10·41%	**7·91%**	12·00%
Warp + MFCC	14·51%	10·48%	9·63%	13·46%
Warp-post + MFCC	14·10%	10·89%	8·34%	13·12%
Warp-post (utt)+ MFCC	13·43%	**9·66%**	8·06%	12·45%
Warp-post + MFCC (joint)	**12·52%**	11·23%	8·98%	**11·98%**

Table 5. *Phone error rate achieved with VTLN approaches to DNN-HMM.*

the time unit used to compute the warping factors. The impact of having a hard or soft decision on the warping factors is studied comparing *Warp + MFCC* to *Warp-post (utt) + MFCC*. While the effects of the time unit used to compute warping factors are studied comparing *Warp-post + MFCC* to *Warp-post (utt) + MFCC*.

On the evaluation set including all the target groups of speakers (ChildIt + APASCI) the VTLN normalisation approach improves the baseline performance by 19% PER relative (from 14.32% to 12.00% PER with $p < .001$). The system working with MFCC features augmented with warping factor improves the baseline by 6% PER relative (from 14.32% to 13.46% PER with $p < .001$). The system working with the MFCC feature vector augmented with the posterior probabilities of the warping factors improves the baseline by 9% relative (from 14.32% to 13.12% PER with $p < .001$) and the system working with the MFCC feature vector augmented with the posterior probabilities of the warping factors averaged at utterance level improves the baseline by 15% relative (from 14.32% to 12.45% PER with $p < .001$). In this latter system however, the averaging operation over utterances of variable length take place between the DNN-warp and the DNN-HMM. Back-propagating the gradient through the variable length averaging in not trivial to implement in practice. Therefore the system *Warp-post (utt)* is not used for joint optimisation. The system performing joint optimisation of the DNN-warp and the DNN-HMM improves the baseline by 19% relative (from 14.32% to 11.98%). The performance differences between the best two system (*VTLN-normalisation* and *Warp-post + MFCC (joint)*) is not statistically significant.

VTLN normalisation allows to consistently obtain PER among the best for each group of speakers. The *Warp-post + MFCC (joint)* overall improvement is mainly due to the large improvement on the children evaluation set, 24% relative (from 15.56% to 12.52% with $p < .001$) whereas it mildly degrades performance on other groups of speakers. This is probably due to the fact that the training set is unbalanced towards children (7h:15m in ChildIt against 2h:40m for each adult group), therefore, performing the joint optimisation biases the system in favour of children's speech.

Using directly warping factors obtained in a standard way (row *Warp + MFCC*), that is augmenting MFCC with a feature vector having a component in correspondence of each possible warping factor with value 1 for the selected warping factor and 0 for all the other warping factors, consistently performs among the worst system and is outperformed by the system using the MFCC augmented with the posterior probabilities of the warping factors. This seems to indicate that the ASR can benefit from the flexibility introduced by the posterior probabilities of the warping factors, in contrast with the hard decision that is the standard warping factors estimation. To perform best however, these estimations have to be conditioned either by averaging at utterance level or by using joint-optimisation. Note that both of these constraints are not compatible in the present framework.

5.1.4 Combination of approaches

Combining several approaches is a common way to improve systems performance and robustness. It is decided here to combine the different approaches introduced up until this point to exploit their potential complementarity. It was chosen to either combine the different approaches at features level (standard VTLN normalised features and the posterior probabilities of the warping factors are combined at the input of the DNN) or to use acoustic features augmented with the posterior probabilities of the warping factors as inputs to a DNN with age-gender adaptation.

Table 6 presents the PER obtained with the DNN-HMM baseline, the age/gender adaptation approach in combination with model selection (row *Model selection*), VTLN approaches (rows *VTLN-normalisation* and *Warp-post + MFCC*) and the combination of the aforementioned approaches: age/gender adaptation performed on a system trained with VTLN-normalised features (row *VTLN (model selection)*), on a system working with the MFCC feature vector augmented with the posterior probabilities of the warping factors (row *Warp-post + MFCC (model selection)*) and on a system trained on VTLN-normalised feature vector augmented with the posterior probabilities of the warping factors (row *Warp-post + VTLN (model selection)*). Joint optimisation is not applied at this stage as the unbalanced training corpus results in biased training and the corpora used here are too small to truncate them to produce a balanced heterogeneous corpus.

On the evaluation set including all the target groups of speakers (ChildIt + APASCI) the combination of approaches outperform all the individual approaches presented until here. The combination *Warp-post + MFCC (model selection)* improves the baseline by 30% relative (from 14.32% PER to 10.98% PER with $p < .001$). *Warp-post + VTLN* improves the baseline by 20% relative (from 14.32% PER to 11.90% PER with $p < .001$) and *VTLN (model selection)* improves the baseline by 35% relative (from 14.32% PER to 10.61% PER with $p < .001$). The combination of the three approaches presented in this paper (*Warp-post + VTLN (model selection)*) improves the baseline by 34% relative (from 14.32% PER to

	Evaluation Set			
	ChildIt	APASCI(f)	APASCI(m)	ChildIt + APASCI
Baseline	15·56%	10·91%	8·62%	14·32%
Model selection	12·43%	9·65%	7·61%	11·59%
Warp-post + MFCC	14·10%	10·89%	8·34%	13·12%
Warp-post + MFCC (model selection)	11·71%	9·23%	7·28%	10·98%
VTLN-normalisation	12·80%	10·41%	7·91%	12·00%
VTLN (model selection)	**11·31%**	9·14%	**7·19%**	**10·61%**
Warp-post + VTLN	**12·64%**	**10·28%**	**8·14%**	**11·90%**
Warp-post + VTLN (model selection)	11·34%	**9·04%**	7·32%	10·68%

Table 6. *Phone error rate achieved with combination of approaches.*

10.68% PER with $p < .001$). The difference between *VTLN (model selection)* and *Warp-post + VTLN (model selection)* is not statistically significant. When compared to the best system until now (*Model selection*), the combination of different approaches improves from 5% relative (*Warp-post + MFCC (model selection)* with $p < .001$) to 9% relative (*VTLN (model selection)* with $p < .001$). The combination *Warp-post + VTLN* on the other hand does not significantly improve the performance compared to *Model selection*. Therefore, this approach will not be considered for further experiments.

The combination of different approaches allows to consistently improve the PER on every group of speakers. On the ChildIt corpus, the best performance are obtained with the system based on VTLN normalised features (*VTLN (model selection)* and *Warp-post + VTLN (model selection)*) which improve by up to 38% PER relative compared to the baseline ($p < .001$) and 10% PER relative ($p < .001$) compared to the best system until now (*Model selection*). On the adult corpora the difference between the performances of the three different combinations of several approaches is not statistically significant. On female speakers, different combination of several approaches allow to improve the baseline by up to 21% PER relative ($p < .001$) and improve the performance of the best system to date (*Model selection*) by up to 7% relative ($p < .01$). On male speakers, different combination of several approaches allow to improve the baseline by up to 20% PER relative ($p < .001$) and improve the performance of the best system to date (*Model selection*) by up

Adaptation	Evaluation Set			
Set	ChildIt	IBN(f)	IBN(m)	ChildIt+IBN
Baseline	12·83%	10·61%	11·02%	11·98%
Model selection	**10·89%**	**10·33%**	**10·99%**	**10·93%**
ChildIt + general model	**10·89%**	10·61%	11·02%	11·00%

Table 7. *Word error rate achieved with the DNN-HMM trained on a mixture of adult and children's speech and adapted to specific age/gender groups.*

to 5% relative ($p < .05$). The combination *Warp-post + MFCC (model selection)* represents the best single-pass decoding system presented here.

5.2 Word recognition

The experiments presented here are designed to verify that results obtained for phone recognition can be replicated in terms of WER and on a more "realistic" set-up where the adult speech training corpus (IBN corpus) is larger than the children speech training corpus (ChildIt). During the experiments the language model weight is tuned on the development set and used to decode the test set. Variations in recognition performance were again validated using the matched-pair sentence test (Gillick and Cox, 1989).

5.2.1 Age/gender adapted DNN-HMM

Table 7 presents the WER obtained with a DNN-HMM baseline trained on the corpus composed of ChildIt and IBN (row *Baseline*). These performance are compared with the performance obtained with age/gender adaptation (row *Model selection*) and with the performance obtained with a system performing model selection between age adapted systems for child speakers and the general baseline for adult speakers (row *ChildIt + general model*).

On the evaluation set including all the target groups of speakers (ChildIt + IBN) the age-gender adaptation improves the performance of the baseline by 10% WER relative (from 11.98% to 10.93% with $p < .001$). When targeting child speakers, the age adaptation improves the performance of the baseline by 18% relative (from 12.83% to 10.89% with $p < .001$). On the other hand, when targeting adult speakers, the age-gender adaptation does not significantly improve the WER compared to the baseline. This is due to the fact that the adult corpus is now considerably larger than for the experiments on PER ($52h : 00m$ for IBN against $5h : 19m$ for APASCI). This allows effective training to be achieved on the adult groups with the general corpus and benefits from age-gender adaptation are limited. Therefore for simplicity's sake, in the remainder of the paper, the approach (row *ChildIt + general model*) is considered instead of age-gender adaptation for all groups of speakers (*Model selection*). The performance difference between *Model selection* and *ChildIt + general model* is not statistically significant.

5.2.2 VTLN based approaches and combination of different approaches

Table 8 presents the WER performance for a) VTLN based approaches: VTLN applied to MFCC during training and testing (row *VTLN-normalisation*), the MFCC features augmented with the posterior probabilities of the warping factors (row *Warp-post + MFCC*) and the joint optimisation of the DNN-warp and the DNN-HMM (row *Warp-post + MFCC (joint)*); b) the combination of several approaches introduced here: VTLN-normalised feature vector augmented with the posterior probabilities of the warping factors and joint optimisation (row *Warp-post + VTLN (joint)*), age adaptation for child speakers performed on a system working with the MFCC feature vector augmented with the posterior probabilities of the warping factors with joint optimisation (row *Warp-post + MFCC (joint/ChildIt + general model)*) and on a system trained on VTLN-normalised feature vector augmented with the posterior probabilities of the warping factors with joint optimisation (row *Warp-post + VTLN (joint/ChildIt + general model)*). These systems are compared to the baseline and to *ChildIt + general model.*

The approach combining VTLN-normalised features and posterior probabilities aims at testing the complementary between VTLN-normalisation that operates at utterances level and posterior probabilities that are obtained at frame level. While estimating VTLN factors on a longer time unit (utterance) should allow for a more accurate average estimation, the "true" warping factor might be fluctuating over time (Miguel, Lieida, Rose, Buera, and Ortega, 2005; Maragakis and Potamianos, 2008). Combining VTLN normalisation at utterance level and posterior probabilities estimated at frame level should help overcoming this problem.

On the evaluation set including all the target groups of speakers (ChildIt + IBN) the VTLN based approaches (*Warp-post + MFCC* and *VTLN-normalisation*) perform similarly (11.57% and 11.58% WER). They improve the performance baseline by 3.5% WER relative ($p < .001$) but both the methods are outperformed by *ChildIt + general model* by 5% WER relative ($p < .001$). The experiments on the children corpus tend to confirm this improvement. Indeed, the systems *Warp-post + MFCC* and *VTLN-normalisation* improve the baseline performance by 6% WER relative (from 12.83% to 12.11% with $p < .001$) and 5% WER relative (from 12.83% to 12.21% with $p < .001$), respectively. Both the approaches are still outperformed on the children corpus by *ChildIt + general model* ($p < .001$). The performance difference between the VTLN based approaches, the baseline and *ChildIt + general model* on adult corpora are in general not statistically significant.

During these experiment, the corpora were unbalanced towards adults ($52h : 00m$ for IBN against $7h : 15m$ for ChildIt). Joint optimisation is performed on a balanced training set in order to avoid introducing a bias in favour of the adult corpora. The balanced corpus is composed of 7h of adult female and 7h of adult male speech

	Evaluation Set			
	ChildIt	IBN(f)	IBN(m)	ChildIt+IBN
Baseline	12·83%	10·61%	11·02%	11·98%
ChildIt + general model	10·89%	10·61%	11·02%	11·00%
Warp-post + MFCC	12.11%	10·52%	11·07%	11·57%
Warp-post + MFCC (joint)	11·81%	**10·49%**	**11·01%**	11·33%
Warp-post + MFCC (joint / ChildIt + general model)	11·06%	**10·49%**	**11·01%**	10·97%
VTLN-normalisation	12·21%	10·58%	11·25%	11·58%
Warp-post + VTLN (joint)	**10·83%**	**10·49%**	11·07%	**10·86%**
Warp-post + VTLN (joint / ChildIt + general model)	11·07%	**10·49%**	11·07%	10·96%

Table 8. *Word error rate achieved with several VTLN approaches to DNN-HMM.*

randomly selected from the IBN corpus. On the evaluation set composed of all target groups, joint optimisation improves the *Warp-post + MFCC* performance by 2% WER relative (from 11.57% to 11.33% with $p < .001$). The performance improvement in each speaker group is not statistically significant.

The combination *Warp-post + MFCC (joint/ChildIt + general model)* improves the *Warp-post + MFCC (joint)* performance by 3% WER relative (from 11.33% to 10.97% $p < .001$). The combination *Warp-post + VTLN (joint)* improves the *VTLN-normalisation* performance by 7% WER relative (from 11.58% to 10.86% $p < .001$). Both these combinations improve the baseline performance by 11% WER relative ($p < .001$). The difference between the three combinations (*Warp-post + MFCC (joint/ChildIt + general model)*, *Warp-post + VTLN (joint)* and *Warp-post + VTLN (joint/ChildIt + general model)*) and the *ChildIt + general model* is not statistically significant. This tendency confirms in each target groups of speakers.

Among the approaches proposed in the paper, *ChildIt + general model* and *Warp-post + VTLN (joint)* perform equally well. However, their potential applications are different. Indeed, *ChildIt + general model* is the most simple approach but lacks flexibility and is difficult to generalise to new groups of speakers as a new DNN would have to be adapted to each new group of speakers. The VTLN based approach *Warp-post + VTLN (joint)* on the other end, does not rely on model adaptation/selection and is more general than *ChildIt + general model.* The drawback of this approach, however, is that it requires a two-pass decoding whereas *ChildIt + general model* operates in a single-pass granted that the age or gender group group is known during decoding.

6 Conclusions

In this paper we have investigated the use of the DNN-HMM approach to speech recognition targeting three groups of speakers, that is children, adult males and adult females. Two different kinds of approaches have been introduced here to cope with inter-speaker variability: approaches based on DNN adaptation and approaches relying on VTLN. The combination of the different approaches to take advantage of their complementarity has then been investigated.

The different approaches presented here have been tested extensively in terms of PER on small corpora first. Systems based on VTLN have been shown to provide a significant improvement compared to the baseline (up to 19% relative) but were still outperformed by the DNN adaptation (23% relative improvement compared to the baseline). The combination of several techniques on the other hand effectively takes advantage of the complementarity of the different approaches introduced in this paper and improves the baseline performance by up to 35% relative PER. Besides, the combination of several techniques is shown to consistently outperform each approach used separately.

Then, the best performing approaches have been validated in terms of WER on a more "realistic" set-up where the adult speech corpus (IBN) used for training is larger than the training children's speech corpus (ChildIt). DNN adaptation is then proved effective for the under-resourced target group (children) but not significantly on the target group with sufficient training data (adults). The trend observed on PER persists and approaches based on VTLN have been shown to provide a significant improvement compared to the baseline (5% to 6% relative) but were still outperformed by the DNN adaptation approach (10% relative improvement compared to the baseline). The combination of different approaches improves the baseline performance by up to 11% WER relative. The two best performing systems introduced here (*ChildIt + general model* and *Warp-post + VTLN (joint)*) perform equally well but can have different applications. Indeed, *ChildIt + general model* is the most simple system but lacks flexibility whereas the VTLN based system *Warp-post + VTLN (joint)* is more general but it requires a two-pass decoding.

Extensions of this work could consider, for comparison or combination with the here proposed approaches, state-of-the-art approaches based on speaker identity models such as I-vectors, speaker codes, linear input networks and linear output networks.

7 Acknowledgements

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

References

O. Abdel-Hamid and H. Jiang. 2013a. Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code. In *Proc of ICASSP*, pages 7942–6.

O. Abdel-Hamid and H. Jiang. 2013b. Rapid and effective speaker adaptation of convolutional neural network based models for speech recognition. In *Proc. of INTERSPEECH*, pages 1248–52.

B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo. 1994. Speaker Independent Continuous Speech Recognition Using an Acoustic-Phonetic Italian Corpus. In *Proc. of ICSLP*, pages 1391–4.

P. Bell, P. Swietojanski, and S. Renals. 2013. Multi-level adaptive networks in tandem and hybrid ASR systems. In *Proc. of ICASSP*, pages 6975–9.

Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. 2007. Greedy layer-wise training of deep networks. *Proc. of NIPS*, 19:153–60.

Y. Bengio, A. Courville, and P. Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–828.

H. A Bourlard and N. Morgan. 1994 *Connectionist speech recognition: a hybrid approach*, volume 247. Springer.

L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, Kai Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiat, D. Povey, A. Rastrow, R.C. Rose, and S. Thomas. 2010. Multilingual acoustic modeling for speech recognition based on subspace Gaussian Mixture Models. In *Proc. of ICASSP*, pages 4334–7.

T. Claes, I. Dologlou, L. ten Bosch, and D. Van Compernolle. 1998. A Novel Feature Transformation for Vocal Tract Length Normalisation in Automatic Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 6(6):549–57.

G.E. Dahl, Dong Yu, Li Deng, and A. Acero. 2012 Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42.

S. Das, D. Nix, and M. Picheny. 1998. Improvements in Children's Speech Recognition Performance. In *Proc. of ICASSP*.

N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. 2011 Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–98.

E. Eide and H. Gish. 1996 A Parametric Approach to Vocal Tract Lenght Normalization. In *Proc. of ICASSP*, pages 346–9.

D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio. 2010 Why does unsupervised pre-training help deep learning? *The Journal of Machine Learning Research*, 11:625–60.

W. T. Fitch and J. Giedd. 1999. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *Journal of Acoustical Society of America*, 106(3):1511–22.

M. J. F. Gales. 1998. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98.

M. Gerosa, D. Giuliani, and F. Brugnara. 2007 Acoustic variability and automatic recognition of children's speech. *Speech Communication*, 49(10–11):847 – 60.

M. Gerosa, D. Giuliani, and F. Brugnara. 2009a. Towards age-independent acoustic modeling. *Speech Communication*, 51(6):499 – 509.

M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos. 2009b. A Review of

ASR Technologies for Children's Speech. In *Proc. of the 2nd Workshop on Child, Computer and Interaction*, pages 1–8.

L. Gillick and S. Cox. 1989. Some Statistical Issues in the Comparison of Speech Recognition Algorithms. In *Proc. of ICASSP*, 1:532–5.

D. Giuliani and M. Gerosa. 2003. Investigating Recognition of Children Speech. In *Proc. of ICASSP*, 2:137–40.

A. Hagen, B. Pellom, and R. Cole. 2003 Children's Speech Recognition with Application to Interactive Books and Tutors. In *Proc. of ASRU.*

G. Hinton, S. Osindero, and Y.-W. Teh. 2006 A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–54.

G. Hinton, Li Deng, Dong Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury. 2012. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, 29(6):82–97.

J. E. Huber, E. T. Stathopoulos, G. M. Curione, T. A. Ash, and K. Johnson. 1999. Formants of children women and men: The effect of vocal intensity variation. *Journal of Acoustical Society of America*, 106(3):1532–42.

D. Imseng, P. Motlicek, P.N. Garner, and H. Bourlard. 2013. Impact of deep MLP architecture on different acoustic modeling techniques for under-resourced speech recognition. In *Proc. of ASRU*, pages 332–7.

Nagendra Kumar and Andreas G. Andreou. 1998. Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. *Speech Communication*, 26(4):283 – 97.

V.-B. Le, L. Lamel, and J. Gauvain. 2010. Multi-style ML features for BN transcription. In *Proc. of ICASSP*, pages 4866–9.

C.-H. Lee and J.-L. Gauvain. 1993. Speaker adaptation based on map estimation of hmm parameters. In *Proc. of ICASSP*, 2:558–61.

L. Lee and R. C. Rose. 1996. Speaker Normalization Using Efficient Frequency Warping Procedure. In *Proc. of ICASSP*, pages 353–6.

S. Lee, A. Potamianos, and S. Narayanan. 1999. Acoustic of children's speech: Developmental changes of temporal and spectral parameters. *Journal of Acoustical Society of America*, 105(3):1455–1468.

B. Li and K.C. Sim. 2010. Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems. In *Proc. of INTERSPEECH*, pages 526–9.

Q. Li and M. Russell. 2001. Why is Automatic Recognition of Children's Speech Difficult? In *Proc. of EUROSPEECH.*

H. Liao. 2013. Speaker adaptation of context dependent deep neural networks. In *Proc. of ICASSP*, pages 7947–51.

M.G. Maragakis and A. Potamianos. 2008. Region-based vocal tract length normalization for ASR. In *Proc. of INTERSPEECH*, pages 1365–8.

A. Metallinou and J. Cheng. 2014. Using Deep Neural Networks to Improve Proficiency Assessment for Children English Language Learners. In *Proc. of INTERSPEECH*, pages 1468–72.

A. Miguel, E. Lieida, R. Rose, L. Buera, and A. Ortega. 2005. Augmented state space acoustic decoding for modeling local variability in speech. In *Proc. of INTERSPEECH*, pages 3009–12.

A. Mohamed, G.E. Dahl, and G. Hinton. 2012. Acoustic Modeling Using Deep Belief Networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 20 (1):14–22.

R. Nisimura, A. Lee, H. Saruwatari, and K. Shikano. 2004. Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability. In *Proc. of ICASSP*.

J. Pinto, Magimai-Doss, and H. Bourlard. 2009. MLP based hierarchical system for task adaptation in ASR. In *Proc. of ASRU*, pages 365–70.

A. Potamianos and S. Narayanan. 2003. Robust Recognition of Children's Speech. *IEEE Transasctions on Speech and Audio Processing*, 11(6):603–15.

G. Saon, H. Soltau, D. Nahamoo, and M. Picheny. 2013. Speaker adaptation of neural network acoustic models using I-vectors. In *Proc. of ASRU*, pages 55–9.

F. Seide, Gang Li, Xie Chen, and Dong Yu. 2011. Feature engineering in Context-Dependent Deep Neural Networks for conversational speech transcription. In *Proc. of ASRU*, pages 24–9.

M. Seltzer, D. Yu, and Y. Wang. 2013. An Investigation Of Deep Neural Networks For Noise Robust Speech Recognition. In *Proc. of ICASSP*.

A. Senior and I. Lopez-Moreno. 2014. Improving DNN speaker independence with I-vector inputs. In *Proc. of ICASSP*.

R. Serizel and D. Giuliani. 2014a. Deep neural network adaptation for children's and adults' speech recognition. In *Proc. of CLIC-It*.

R. Serizel and D. Giuliani. 2014b. Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. In *Proc. of SLT*.

S. Sivadas and H. Hermansky. 2004. On use of task independent training data in tandem feature extraction. In *Proc. of ICASSP*, 1:541–4.

S. Steidl, G. Stemmer, C. Hacker, E. Nöth, and H. Niemann. 2003. Improving Children's Speech Recognition by HMM Interpolation with an Adults' Speech Recognizer. In *Pattern Recognition, 25th DAGM Symposium*, pages 600–7.

A. Stolcke, F. Grezl, Mei-Yuh Hwang, Xin Lei, N. Morgan, and D. Vergyri. 2006. Cross-Domain and Cross-Language Portability of Acoustic Features Estimated by Multilayer Perceptrons. In *Proc. of ICASSP*, 1:321–34.

P. Swietojanski, A. Ghoshal, and S. Renals. 2012. Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR. In *Proc. of SLT*, pages 246–51.

S. Thomas, M.L. Seltzer, K. Church, and H. Hermansky. 2013. Deep neural network features and semi-supervised training for low resource speech recognition. In *Proc. of ICASSP*, pages 6704–8.

Karel Veselỳ, Lukáš Burget, and František Grézl. 2010. Parallel training of neural networks for speech recognition. In *Text, Speech and Dialogue*, pages 439–46. Springer.

S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin. 1996. Speaker Normalisation on Conversational Telephone Speech. In *Proc. of ICASSP*, 1:339–41.

L. Welling, S. Kanthak, and H. Ney. 1999. Improved Methods for Vocal Tract Normalization. In *Proc. of ICASSP*, 2:761–4.

J. G. Wilpon and C. N. Jacobsen. 1996. A Study of Speech Recognition for Children and Elderly. In *Proc. of ICASSP*, pages 349–52.

M. Wöllmer, B. Schuller, A. Batliner, S. Steidl, and D. Seppi. 2011. Tandem Decoding of Children's Speech for Keyword Detection in a Child-robot Interaction Scenario. *ACM Transasctions Speech Language Processing*, 7(4):12:1–12:22.

P.C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young. 1994. Large vocabulary continuous speech recognition using HTK. In *Proc. of ICASSP*, 2:125–8.

K. Yochai and N. Morgan. 1992. GDNN: a gender-dependent neural network for continuous speech recognition. In *Proc. of Iternational Joint Conference on Neural Networks*, 2:332–7.