Natural Language Engineering **23** (3): 325–350. © Cambridge University Press 2016
doi:10.1017/S135132491600005X

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

ROMAIN SERIZEL[1,2] and DIEGO GIULIANI[2]

[1]*LTCI, CNRS, Télécom ParisTech, Université Paris - Saclay*
46, rue Barrault, Paris, 75013, France
e-mail: romain.serizel@telecom-paristech.fr
[2]*HLT research unit, Fondazione Bruno Kessler (FBK)*
Via Sommarive 18, Trento, 38121, Italy
e-mail: giuliani@fbk.eu

(*Received 20 March 2015; revised 29 February 2016; accepted 1 March 2016; first published online 12 April 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 normalised with standard VTLN or mel frequency cepstral coefficients 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 thirty per cent to thirty-five per cent relative and the baseline word error rate performance by about ten per cent relative.

1 Introduction

Speaker-related acoustic variability is a major source of errors in automatic speech recognition (ASR). 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.

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

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 *et al.* 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 years, 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 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 *et al.* 1998; Das, Nix and Picheny 1998; Li and Russell 2001; Giuliani and Gerosa 2003; Potamianos and Narayanan 2003; Gerosa, Giuliani and Brugnara 2007; Gerosa *et al.* 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 *et al.* 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 *et al.* 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 *et al.* 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 *et al.* 2007; Erhan *et al.* 2010; Seide *et al.* 2011), DNN has proven to be an effective alternative to Gaussian mixture model (GMM) in HMM–GMM-based ASR (Bourlard and Morgan 1994; Hinton *et al.* 2012) and really good performance has been obtained with hybrid DNN–HMM systems (Dahl *et al.* 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 *et al.* 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 *et al.* 2013) and the task specific corpus is used for the final training/adaptation of the DNN. In underresourced scenarios, approaches based on DNN (Imseng *et al.* 2013) have then shown to outperform approaches based on subspace GMM (Burget *et al.* 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 *et al.* (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 underresourced 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 unnormalised 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 *et al.* 1996) can help to improve the performance. Here, we consider first the application of a conventional VTLN technique to normalise mel

frequency cepstral coefficients (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 *et al.* 2011; Saon *et al.* 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.

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

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

$$h = \mathbf{w}.\mathbf{y} + b \tag{1}$$

$$\sigma(h) = \frac{1}{1 + e^{-h}} \tag{2}$$

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:

$$softmax(h_j) = \frac{e^{h_j}}{\sum_i e^{h_i}} \tag{3}$$

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:

$$p(X|S) \propto \frac{p(S|X)}{p(S)}, \tag{4}$$

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 (1,700 hours in Senior and Lopez-Moreno (2014)). However, this solution does not apply when tackling ASR for underresourced 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

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

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 underresourced 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 unnormalised 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 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 maximising the likelihood of warped data given its transcription and a current set of HMM-based acoustic models (Lee and Rose

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

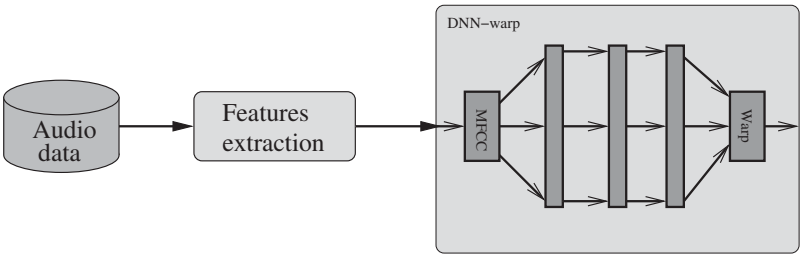

Fig. 1. Training of the DNN-warp.

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 estimated posterior probabilities are appended to the unnormalised 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;

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

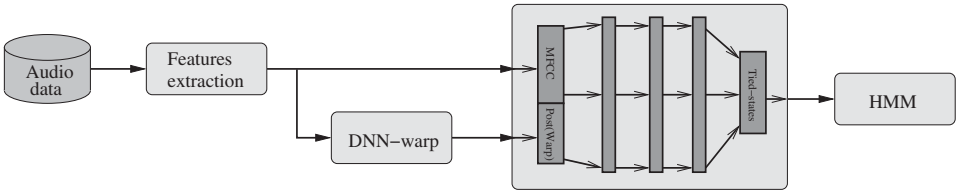

Fig. 2. Training of the warping factor aware DNN–HMM.

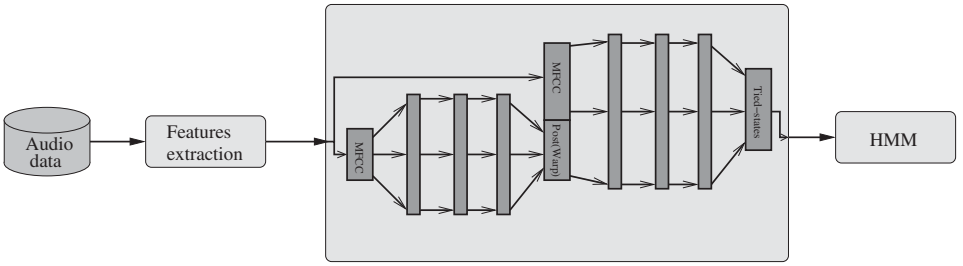

Fig. 3. Joint optimisation of the DNN-warp and the DNN–HMM.

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.

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 five times as much training data as ChildIt or APASCI (Table 1).

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

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

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

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*

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

4.1.1 ChildIt

The ChildIt corpus (Giuliani and Gerosa 2003; Gerosa *et al.* 2007) consists of Italian read sentences collected from 171 children (eighty-six male and eighty-five female) aged between 7 and 13 years, 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 10 h:24 m. 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 7 h:15 m; a development set consisting of data from fourteen speakers, for a total durations of 0 h:49 m; a test set consisting of data from forty-two speakers balanced with respect to age and gender for a total duration of 2 h:20 m.

4.1.2 APASCI

The APASCI speech corpus (Angelini *et al.* 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 7 h:05 m. 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 5 h:19 m; a development set consisting of data from thirty speakers balanced per gender, for a total durations of 0 h:39 m; a test set consisting of data from thirty speakers balanced per gender, for a total duration of 0 h:40 m.

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

4.1.3 IBN corpus

The IBN corpus is composed of speech from several radio and television Italian news programmes (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 52 h:00 m of speech, and a test set formed by 2 h:00 m of speech. During the experiments presented here, 2 h:00 m of male speech and 2 h:00 m 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 25 h:00 m of male speech and 23 h:00 m 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 thirteen MFCC, including the zero-order coefficient, computed on 20 ms frames with 10 ms overlap. First, second- and third-order time derivatives are computed after cepstral mean subtraction performed utterance by utterance. These features are arranged into a fifty-two-dimensional vector that is projected into a thirty-nine-dimensional feature space by applying a linear transformation estimated through Heteroscedastic Linear Discriminant Analysis (Kumar and Andreou 1998).

Acoustic models are 3,039 tied-state triphone HMMs based on a set of forty-eight phonetic units derived from the SAMPA Italian alphabet. Each tied-state is modelled with a mixture of eight Gaussian densities having a diagonal covariance matrix. In addition, 'silence' is modelled with a GMM having thirty-two Gaussian densities.

4.2.2 DNN–HMM

The DNN uses again thirteen MFCC, including the zero-order coefficient, computed on 20 ms frames with 10 ms overlap. The context spans on a thirty-one frames window. For each frequency band, the thirty-one coefficients context is separately scaled with a Hamming window and projected to a sixteen-dimensional vector using DCT. The thirteen 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 3,039 tied-states

obtained from the GMM–HMM training on the mixture of adults' and children's speech (ChildIt + APASCI). The DNN has four hidden layers, each of which contains 1,500 elements such that the DNN architecture can be summarised as follows: 208 × 1,500 × 1,500 × 1,500 × 1,500 × 3,039.

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 ten iterations with a learning rate of 0.005. The following layers are pre-trained with a Bernouilli–Bernouilli RBM trained during five 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 per cent 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 per cent. 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 six hidden layers will result in lower performance than with four 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 forty-eight to twenty-eight 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

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

models are 5,021 tied-state triphone HMM based on a set of forty-eight phonetic units derived from the SAMPA Italian alphabet. Each tied-state is modelled with a mixture of thirty-two Gaussian densities having a diagonal covariance matrix. In addition, 'silence' is modelled with a GMM having thirty-two 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 5,021 tied-states obtained from the word recognition GMM–HMM training on the mixture of adults' and children's speech (ChildIt + IBN). The DNN has four hidden layers, each of which contains 1,500 elements such that the DNN architecture can be summarised as follows: $208 \times 1{,}500 \times 1{,}500 \times 1{,}500 \times 1{,}500 \times 5{,}021$.

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 1.6 G words. Part of the textual data, consisting in about 1.0 G words, were acquired via web crawling of news domains. The recognition dictionary consists of the most frequent 250 K 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 twenty-five 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 twenty-five 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 unwarped children's and adults' speech (Gerosa *et al.* 2007, 2009a).

The DNN-warp inputs are the MFCC with a sixty-one frames context window, DCT projected to a 208 dimensional feature vector (the procedure is similar as in 4.2.2). The targets are the twenty-five warping factors. The DNN has four hidden

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

layers, each of which contains 500 elements such that the DNN architecture can be summarised as follows: $208 \times 500 \times 500 \times 500 \times 500 \times 25$. 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 four hidden layers, each of which contains 1,500 elements such that the DNN architecture can then be summarised as follows: $233 \times 1{,}500 \times 1{,}500 \times 1{,}500 \times 1{,}500 \times 3{,}039$ for phone recognition and $233 \times 1{,}500 \times 1{,}500 \times 1{,}500 \times 1{,}500 \times 5{,}021$ 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 four hidden layers (corresponding to the DNN-warp) and to 0.0001 in the last four 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 four hidden layers and the last four hidden layers is done in an attempt to overcome the vanishing gradient effect in the eight 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, especially 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 altern-

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

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

	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%

ative 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 twenty-eight 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 0.05, 0.01 and 0.001.

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 7 h of training data which is apparently sufficient to train an effective DNN and we can observe an improvement of twenty-two per cent PER relative compared to the baseline performance (from 15.56 per cent to 12.76 per cent with $p < 0.001$). However, in adult data, there is only about 2 h:40 m 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 fourteen per cent PER relative on female speakers in APASCI (from 10.91 per cent to 12.75 per cent with $p < 0.001$) and twelve per cent PER relative on male speakers in APASCI (from 8.62 per cent to 9.83 per cent with $p < 0.001$).

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

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*

	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%

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 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 twenty-five per cent PER relative (from 15.56 per cent to 12.43 per cent with $p < 0.001$). On adult female speakers in APASCI the age/gender adaptation improves the baseline performance by about thirteen per cent PER relative (from 10.91 per cent to 9.65 per cent with $p < 0.001$). On adult male speakers, the age/gender adaptation improves the baseline performance by thirteen per cent (from 8.62 per cent to 7.61 per cent with $p < 0.05$).

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 twenty-three per cent PER relative (from 14.32 per cent to 11.59 per cent with $p < 0.05$).

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

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

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

	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%**

highest likelihood is retained. Comparing recognition results in the last two rows of Table 4, it is possible to note that the automatic model selection approach results in an overall decrease of performance: from 11.59 per cent to 12.26 per cent 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 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 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 nineteen per cent PER relative (from 14.32 per cent to 12.00 per cent PER

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

with $p < 0.001$). The system working with MFCC features augmented with warping factor improves the baseline by six per cent PER relative (from 14.32 per cent to 13.46 per cent PER with $p < 0.001$). The system working with the MFCC feature vector augmented with the posterior probabilities of the warping factors improves the baseline by nine per cent relative (from 14.32 per cent to 13.12 per cent PER with $p < 0.001$) 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 fifteen per cent relative (from 14.32 per cent to 12.45 per cent PER with $p < 0.001$). 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 nineteen per cent relative (from 14.32 per cent to 11.98 per cent). 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, twenty-four per cent relative (from 15.56 per cent to 12.52 per cent with $p < 0.001$), 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 (7 h:15 m in ChildIt against 2 h:40 m 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.

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

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

	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 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 thirty per cent relative (from 14.32 per cent PER to 10.98 per cent PER with $p < 0.001$). *Warp-post + VTLN* improves the baseline by twenty per cent relative (from 14.32 per cent PER to 11.90 per cent PER with $p < 0.001$) and *VTLN (model selection)* improves the baseline by thirty-five per cent relative (from 14.32 per cent PER to 10.61 per cent PER with $p < 0.001$). The combination of the three approaches presented in this paper (*Warp-post + VTLN (model selection)*) improves the baseline by thirty-four per cent relative (from 14.32 per cent PER to 10.68 per cent PER with $p < 0.001$). 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 five per cent relative (*Warp-post + MFCC (model selection)* with $p < 0.001$) to nine per cent relative (*VTLN (model selection)* with $p < 0.001$). The combination *Warp-post + VTLN* on the other hand does not

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

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 thirty-eight per cent PER relative compared to the baseline ($p < 0.001$) and ten per cent PER relative ($p < 0.001$) 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 twenty-one PER relative ($p < 0.001$) and improve the performance of the best system to date (*Model selection*) by up to seven per cent relative ($p < 0.01$). On male speakers, different combination of several approaches allow to improve the baseline by up to twenty per cent PER relative ($p < 0.001$) and improve the performance of the best system to date (*Model selection*) by up to five per cent relative ($p < 0.05$). 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 ten per cent WER relative (from 11.98 per cent to 10.93 per cent with $p < 0.001$). When targeting child speakers, the age adaptation improves the performance of the baseline by eighteen per cent relative (from 12.83 per cent to 10.89 per cent with $p < 0.001$). 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

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

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*

Adaptation set	Evaluation 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%

(52 h : 00 m for IBN against 5 h : 19 m 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 *et al.* 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 per cent and 11.58 per cent WER). They improve the

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

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

	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%

performance baseline by 3.5 per cent WER relative ($p < 0.001$) but both the methods are outperformed by *ChildIt + general model* by five per cent WER relative ($p < 0.001$). 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 six per cent WER relative (from 12.83 per cent to 12.11 per cent with $p < 0.001$) and five per cent WER relative (from 12.83 per cent to 12.21 per cent with $p < 0.001$), respectively. Both the approaches are still outperformed on the children corpus by *ChildIt + general model* ($p < 0.001$). 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 (52 h : 00 m for IBN against 7 h : 15 m 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 7 h of adult female and 7 h of adult male speech randomly selected from the IBN corpus. On the evaluation set composed of all target groups, joint optimisation improves the *Warp-post + MFCC* performance by two per cent WER relative (from 11.57 per cent to 11.33 per cent with $p < 0.001$). 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 three per cent WER relative (from 11.33 per cent to 10.97 per cent $p < 0.001$). The combination *Warp-post + VTLN (joint)* improves the *VTLN-normalisation* performance by seven per cent WER relative (from 11.58 per cent to 10.86 per cent $p < 0.001$). Both these combinations improve the baseline performance by eleven per cent WER relative ($p < 0.001$). 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.

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

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 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 nineteen per cent relative) but were still outperformed by the DNN adaptation (twenty-three per cent 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 thirty-five per cent 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 underresourced 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 (five per cent to six per cent relative) but were still outperformed by the DNN adaptation approach (ten per cent relative improvement compared to the baseline). The combination of different approaches improves the baseline performance by up to eleven per cent 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

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

models such as i-vectors, speaker codes, linear input networks and linear output networks.

References

Abdel-Hamid, O., and Jiang, H. 2013a. Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code. In *Proceedings of ICASSP*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), pp. 7942–6.

Abdel-Hamid, O., and Jiang, H. 2013b. Rapid and effective speaker adaptation of convolutional neural network based models for speech recognition. In *Proceedings of INTERSPEECH*, ISCA, Grenoble, France (Interspeech, ICSLP, Workshop on Child, Computer and Interaction), pp. 1248–52.

Angelini, B., Brugnara, F., Falavigna, D., Giuliani, D., Gretter, R., and Omologo, M. 1994. Speaker independent continuous speech recognition using an acoustic-phonetic Italian corpus. In *Proceedings of ICSLP*, ISCA, Grenoble, France (Interspeech, ICSLP, Workshop on Child, Computer and Interaction), pp. 1391–4.

Bell, P., Swietojanski, P., and Renals, S. 2013. Multi-level adaptive networks in tandem and hybrid ASR systems. In *Proceedings of ICASSP*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), pp. 6975–9.

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

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. 2007. Greedy layer-wise training of deep networks. In *Proceedings of NIPS*, NIPS, La Jolla (CA), United States, vol. 19, 153–60.

Bourlard, H., and Morgan, N. 1994. *Connectionist Speech Recognition: A Hybrid Approach*, Springer, Berlin, Germany, vol. 247, Springer.

Burget, L., Schwarz, P., Agarwal, M., Akyazi, P., Feng, K., Ghoshal, A., Glembek, O., Goel, N., Karafiat, M., Povey, D., Rastrow, A., Rose, R. C., and Thomas, S. 2010. Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models. In *Proceedings of ICASSP*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), pp. 4334–7.

Claes, T., Dologlou, I., ten Bosch, L., and Van Compernolle, D. 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.

Dahl, G. E., Yu, D., Deng, L., and Acero, A. 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.

Das, S., Nix, D., and Picheny, M. 1998. Improvements in Children's speech recognition performance. In *Proceedings of ICASSP*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN).

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

Eide, E., and Gish, H. 1996. A parametric approach to vocal tract lenght normalization. In *Proceedings of ICASSP*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), pp. 346–9.

Erhan, D., Bengio, Y., Courville, A., Manzagol, P.-A., Vincent, P., and Bengio, S. 2010. Why does unsupervised pre-training help deep learning? *The Journal of Machine Learning Research* **11**(2): 625–60.

https://doi.org/10.1017/S135132491600005X Published online by Cambridge University Press

Fitch, W. T., and Giedd, J. 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.

Gales, M. J. F. 1998. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language* **12**(2): 75–98.

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

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

Gerosa, M., Giuliani, D., Narayanan, S., and Potamianos, A. 2009b. A Review of ASR technologies for Children's speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, ISCA, Grenoble, France (Interspeech, ICSLP, Workshop on Child, Computer and Interaction), pp. 1–8.

Gillick, L., and S. Cox, S. 1989. Some statistical issues in the comparison of speech recognition algorithms. In *Proceedings of ICASSP*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), **1**: 532–5.

Giuliani, D., and Gerosa, M. 2003. Investigating recognition of children speech. In *Proceedings of ICASSP* IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), **2**: 137–40.

Hagen, A., Pellom, B., and Cole, R. 2003. Children's speech recognition with application to interactive books and tutors. In *Proceedings of ASRU*. IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN).

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. 2012. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine* **29**(6): 82–97.

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

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

Imseng, D., Motlicek, P., Garner, P. N., and Bourlard, H. 2013. Impact of deep MLP architecture on different acoustic modeling techniques for under-resourced speech recognition. In *Proceedings of ASRU*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), pp. 332–7.

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

Le, V.-B., Lamel, L., and Gauvain, J.-L. 2010. Multi-style ML features for BN transcription. In *Proceedings of ICASSP*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), pp. 4866–9.

Lee, C.-H., and Gauvain, J.-L. 1993. Speaker adaptation based on map estimation of hmm parameters. In *Proceedings of ICASSP*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), vol. 2, 558–61.

Lee, L., and Rose, R. C. 1996. Speaker normalization using efficient frequency warping procedure. In *Proceedings of ICASSP*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), pp. 353–6.

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

Li, B., and Sim, K. C. 2010. Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems. In *Proceedings of INTERSPEECH*, ISCA, Grenoble, France (Interspeech, ICSLP, Workshop on Child, Computer and Interaction), pp. 526–9.

Li, Q., and Russell, M. 2001. Why is automatic recognition of children's speech difficult? In *Proceedings of EUROSPEECH*. ISCA, Grenoble, France (Interspeech, ICSLP, Workshop on Child, Computer and Interaction).

Liao, H. 2013. Speaker adaptation of context dependent deep neural networks. In *Proceedings of ICASSP*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), pp. 7947–51.

Maragakis, M. G., and A. Potamianos, A. 2008. Region-based vocal tract length normalization for ASR. In *Proceedings of INTERSPEECH*, ISCA, Grenoble, France (Interspeech, ICSLP, Workshop on Child, Computer and Interaction), pp. 1365–8.

Metallinou, A., and Cheng, J. 2014. Using deep neural networks to improve proficiency assessment for children english language learners. In *Proceedings of INTERSPEECH*, ISCA, Grenoble, France (Interspeech, ICSLP, Workshop on Child, Computer and Interaction), pp. 1468–72.

Miguel, A., Lieida, E., Rose, R., Buera, L., and Ortega, A. 2005. Augmented state space acoustic decoding for modeling local variability in speech. In *Proceedings of INTERSPEECH*, ISCA, Grenoble, France (Interspeech, ICSLP, Workshop on Child, Computer and Interaction), pp. 3009–12.

Mohamed, A., Dahl, G. E., and Hinton, G. 2012. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech and Language Processing* **20**(1): 14–22.

Nisimura, R., Lee, A., Saruwatari, H., and Shikano, K. 2004. Public speech-oriented guidance system with adult and child discrimination capability. In *Proceedings of ICASSP*. IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN).

Pinto, J., Magimai-Doss, N., and Bourlard, H. 2009. MLP based hierarchical system for task adaptation in ASR. In *Proceedings of ASRU*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), pp. 365–70.

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

Saon, G., Soltau, H., Nahamoo, D., and Picheny, M. 2013. Speaker adaptation of neural network acoustic models using I-vectors. In *Proceedings of ASRU*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), pp. 55–9.

Seide, F., Li, G., Chen, X., and Yu, D. 2011. Feature engineering in context-dependent deep neural networks for conversational speech transcription. In *Proceedings of ASRU*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), pp. 24–9.

Seltzer, M. L., Yu, D., and Wang, Y. 2013. An investigation of deep neural networks for noise robust speech recognition. In *Proceedings of ICASSP*. IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN).

Senior, A., and Lopez-Moreno, I. 2014. Improving DNN speaker independence with I-vector inputs. In *Proceedings of ICASSP*. IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN).

Serizel, R., and Giuliani, D. 2014a. Deep neural network adaptation for children's and adults' speech recognition. In *Proceedings of CLIC-It*. Pisa University Press, Pisa, Italy (CLIC-It).

Serizel, R., and Giuliani, D. 2014b. Vocal tract length normalisation approaches to DNN-based children's and adults' speech recognition. In *Proceedings of SLT*. IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN).

Sivadas, S., and Hermansky, H. 2004. On use of task independent training data in tandem feature extraction. In *Proceedings of ICASSP* vol. 1, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), pp. 541–4.

Steidl, S., Stemmer, G., Hacker, C., Nöth, E., and Niemann, H. 2003. Improving children's speech recognition by HMM interpolation with an adults' speech recognizer. In *Proceedings of Pattern Recognition, 25th DAGM Symposium*, Springer, Berlin, Germany (Pattern Recognition - 25th DAGM Symposium), pp. 600–7.

Stolcke, A., Grézl, F., Hwang, M.-Y., Lei, X., Morgan, N., and Vergyri, D. 2006. Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons. In *Proceedings of ICASSP*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), vol. 1, pp. 321–34.

Swietojanski, P., Ghoshal, A., and Renals, S. 2012. Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR. In *Proceedings of SLT*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), pp. 246–51.

Thomas, S., Seltzer, M. L., Church, K., and H. Hermansky, H. 2013. Deep neural network features and semi-supervised training for low resource speech recognition. In *Proceedings of ICASSP*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), pp. 6704–8.

Veselỳ, K., Burget, L., and Grézl, F. 2010. Parallel training of neural networks for speech recognition. In *Text, Speech and Dialogue*, Springer, Berlin, Germany, pp. 439–46. Springer.

Wegmann, S., McAllaster, D., Orloff, J., and Peskin, B. 1996. Speaker normalisation on conversational telephone speech. In *Proceedings of ICASSP*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), vol. 1, 339–41.

Welling, L., Kanthak, S., and Ney, H. 1999. Improved methods for vocal tract normalization. In *Proceedings of ICASSP*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), vol. 2, 761–4.

Wilpon, J. G., and Jacobsen, C. N. 1996. A study of speech recognition for children and elderly. In *Proceedings of ICASSP*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), pp. 349–52.

Wöllmer, M., Schuller, B., Batliner, A., Steidl, S., and Seppi, D. 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.

Woodland, P. C., Odell, J. J., Valtchev, V., and Young, S. J. 1994. Large vocabulary continuous speech recognition using HTK. In *Proceedings of ICASSP*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), vol. 2, pp. 125–8.

Yochai, K., and Morgan, N. 1992. GDNN: a gender-dependent neural network for continuous speech recognition. In *Proceedings of Iternational Joint Conference on Neural Networks*, IEEE, New York (NY), United States (ICASSP, ASRU, SLT, IJCNN), vol. 2, pp. 332–7.