In this paper, DNN is used to deal with the acoustic variability induced by age and gender differences in heterogeneous groups of speakers including children for speech recognition. Both speaker normalization and adaptation techniques are used.

Speaker normalization attempt 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. This process is known as vocal tract length normalization (VTLN). This paper has evaluated two schemes of incorporating VTLN for DNN-based speech recognition systems. The first approach does a grid search of the VTLN warp factors via maximum likelihood using the standard GMM-HMM based system. The warp factors are then used to extract the MFCC features to train a standard DNN-HMM system. Alternatively, an additional DNN is trained to predict the pre-defined set of 25 warping factors. The warping factor posteriors are then augmented to the standard MFCC features to train the standard DNN-HMM system for the recognition task. Finally, the DNN used to predict the warp factors and the backend DNN can be jointly optimized via standard error back-propagation.

For speaker adaptation, a single age/gender-independent DNN is trained with all the available training data. It is then “adapted” using the data from the age/gender specific corpus.

In general, the paper is well-written and easy to follow. The problem to be addressed is also well-defined. The idea of jointly optimizing two neural networks is also interesting.

However, one of my biggest concerns is that the paper lacks technical novelties. In addition, the experiments are not comprehensive enough to compare with the state-of-the-art approaches. The following are some of my comments:

1. In section 4.5, the authors conduct a grid search based on a set of 25 warping factors via the standard maximum likelihood using a triphone single Gaussian system. My question is that are there any difficulties in using a more robust triphone GMM system to search the warping factors? Single Gaussian is definitely not robust enough for a reliable warping factor. The easiest argument is that at least two Gaussians are needed to model both genders. In addition, what if the warping factor is per speaker rather than per utterance?
2. The speaker adaptation is achieved by retraining of an age/gender independent DNN with an age/gender specific corpus “adaptation”. How does this approach compare with the state-of-the-art DNN adaptation schemes? For example, using

ivectors or speaker codes as additional inputs to the DNN? What if the speaker-adapted features (e.g., fMLLR) are used as the DNN inputs?

3. How is system combination performed? I am not sure I understand what the authors mean by “features level” in section 5.1.4. More implementation details are appreciated. How about the recognition-level combination schemes, e.g., confusion network combination? Bayesian risk decoding? etc.