# Vocal tract length compensation in the signal and model domains in child speech recognition

*Mats Blomberg and Daniel Elenius*
*Department of Speech, Music and Hearing, CSC, KTH, Stockholm*

## *Abstract*

*In a newly started project, KOBRA, we study methods to reduce the required amount of training data for speech recognition by combining the conventional data-driven training approach with available partial knowledge on speech production, implemented as transformation functions in the acoustic, articulatory and speaker characteristic domains. Initially, we investigate one well-known dependence, the inverse proportional relation between vocal tract length and formant frequencies. In this report, we have replaced the conventional technique of frequency warping the unknown input utterance (VTLN) by transforming the training data instead. This enables phoneme-dependent warping to be performed. In another experiment, we expanded the available training data by duplicating each training utterance into a number of differently warped instances. Training on this expanded corpus results in models, each one representing the whole range of vocal tract length variation. This technique allows every frame of the utterance to be warped differently. The computational load is reduced by an order of magnitude compared to conventional VTLN without noticeable decrease in performance on the task of recognising children's speech using models trained on adult speech.*

## Introduction

Speech recognition has reached a performance level which has enabled implementation in many practical applications. However, although the systems often function well for most people and in most conditions, there are still many users who experience an uncomfortably high frequency of recognition errors.

The background of this paper is our opinion that an explanation to limited robustness and accuracy of speech recognition systems is often to be found in the standard procedures of training and adapting the acoustic models. The current strategy to record all possible combinations of variability in speaker characteristics, speaking styles, and environment conditions during training leads, ultimately, to non-realistic size requirements on the training data (Moore, 2003). We have to find new ways of improving speech recognition performance.

In a newly started project, KOBRA (KnOwledge-Based and –Rich Adaptation), we study methods for reducing the required amount of training data. The basic idea is to combine knowledge of acoustic-phonetic properties and relations with conventionally trained models in order to predict speech from speaker categories not present in the training data. If successful, this technique could reduce the size requirement by modelling speech variability using existing knowledge rather than learning on training data.

In this paper we exemplify the approach by reporting results from experiments on recognition of children's speech using models trained on adult speech. We choose a prominent speaker characteristic feature, vocal tract size for predicting child models by transforming the adult training speech data. The technique is compared with the standard Vocal Tract Length Normalization (VTLN) technique.

## Approach

### Vocal Tract Length Normalization

In vocal tract length normalization (Lee and Rose, 1996), the aim is to suppress the implications of the vocal tract length of a speaker and hence disregard spectral dissimilarities thereof. This can be accomplished by performing a frequency axis expansion or compression of the spectrum. After this warping, there should be better correspondence between formant positions of an unknown speaker and those of the training data.

In VTLN a challenge is to estimate the amount of warping which results in optimal recognition performance. One method to estimate such a warping factor is based on the ability of a statistically based recognizer to rank utterances based on the likelihood of having been generated by it's acoustic model. An assumption is made that the warping resulting in a high acoustic likelihood also results in a high recognition accuracy.

During recognition, the test utterance is sequentially warped with each warp factor from a pre-determined set of factor values. A recognition phase is performed for each value and among the set of warping factor specific recogniser outputs, that identity is selected which has the highest matching score.

A grid search to find an optimal warp-factor results in a high computational burden. This burden can be lowered by for example reducing the complexity of the acoustic model used in warp estimation, as in (Welling, Kanthak and Ney, 1999). Thereby less computations are needed and time is saved.

### Phoneme-dependent warping

Traditional VTLN is constrained to perform uniform warping of the whole input utterance since the phoneme identities and time boundaries are not known before the utterance is recognised. This is not appropriate for those production components which influence different phonemes in different manners. In order to perform phone-dependent warping a change of domain is necessary.

Phoneme-dependent warping can be implemented by adapting the acoustic models of a phone-based recognition system. Thereby manipulation of separate phonemes is possible. Two means to accomplish Vocal Tract Length Adaptation (VTLA) are 1) direct transformation of the trained model parameters and 2) indirect transformation by applying signal processing on the training data.

Model parameter transformation is in general faster than training models from scratch on transformed training data. The resolution is however limited by the model parameters used. Direct transformation of model parameters such as variance, velocity and acceleration coefficients also requires special care.

Voice transformation, e.g. frequency warping, of the training speech data can be used to alter the speech quality prior to training a statistical model. This allows detailed manipulation of individual utterances but two problems with this technique is that the training involved for a set of warp factors is substantial when the training corpora size is large and warping can result in spectral artefacts which restricts the clarity of the models. No alteration of the standard training procedure is needed. The only difference is that the set of training data is warped prior to training. Thereby it is possible to produce a phone model superset, in which each phone has a number of separate model versions, one for each warp factor.

This paper will focus on the first approach, application of VTLN on the training data. The objective is to verify that the method works in practice.

### Phoneme-dependent warp estimation

As in VTLN, the optimal degree of warping in VTLA is estimated by maximising the matching score. It is theoretically possible to proceed as in VTLN by an exhaustive search for the set of differently warped models that results in the highest score for a given test utterance. However, this is not feasable since the search space is huge. Even a simple example with only two possible warping factors and a restricted set of 20 phones results in over a million different model combinations, each one having to be evaluated by a complete recognition procedure of the utterance.

Two methods to reduce the search space have been studied in this paper. One technique is based on the assumption that score maximisation can be performed independently for each phoneme. The second technique avoids explicit warping altogether using a *multi-warp* model which is trained on the complete set of transformed training speech for all warping factors.

## Recognition task

Performance was measured using a connected digit-task to keep other sources of variation low. Co-articulation variation due to phoneme-context was limited by the small vocabulary size and the language model was kept uniform by using random digit-strings.

## Corpora

The experiments were conducted on Swedish speakers recorded in the Pf-Star (Batliner, Blomberg, D'Acry, Elenius and Giuliani, 2005) and SpeeCon (Großkopf, Marasek, Heuvel, Diehl and Kiessling, 2002) corpora. Both contain prompted digit-strings recorded using a head-set with a directional Sennheiser ME 104 microphone.

SpeeCon consists of both adults and children down to the age of 8 years (Großkopf et.al. 2002). In these recordings the speakers read all prompts from a computer screen. Recordings were collected in an office environment. Analog to digital conversion was performed using 16 bits at 16 kHz. The number of digits per speaker was 30, divided into a 10-digit string and four 5-digit strings.

The Swedish part of Pf-Star consists of 198 children of 4 to 8 years repeating oral prompts spoken by an adult speaker (Batliner et.al. 2005). Only connected-digit strings were used in this study to concentrate on acoustic modelling rather than language models. Each child was prompted to speak 10 three-digit strings amounting to 30 digits. Recordings were made in a separate room at day care and after-school centres. Analog to digital conversion was specified to be 24 bits / 32 kHz. However, due to an error in the hardware equipment or the interface, the actual sampling frequency was 7% higher, 34.2 kHz. A compensation was performed in January 2007 by correcting the stored sampling frequency value in the information header of all sound files. Down sampling was then performed to 16 bits / 16 kHz to match that of SpeeCon.

# Recognition System

In the connected digit-string grammar, each utterance was assumed to begin and end with a silent segment framing a block of an arbitrary number of digits with optional short silence intervals in between. Words were modeled using context independent phoneme models.

The temporal aspect of a phone was modeled using a three-state Hidden Markov Model (HMM). In each state the acoustic feature vector was modeled using Gaussian Mixtures with a diagonal covariance matrix.

Speech spectra were represented by a 39 element feature vector. These coefficients consisted of 13 MFCC (Mel Frequency Cepstrum Coefficients) and their first and second order time derivatives (velocity and acceleration).

Feature extraction involved dividing the speech signal into 10 ms frames using a 28 ms Hamming window. The MFCCs were calculated using a cosine transform of the output from a mel-scaled filterbank consisting of 26 channels in the interval 0 to 7.6 kHz.

Recognition models were initiated and trained similarly to the RefRec script (Lindberg et.al., 2000). All training and recognition experiments were implemented in the HTK speech recognition software package (Young et.al., 2005).

# Experiment

An initial experiment was performed to evaluate the concept of extending the training data based on vocal tract size.

The adult training data was extended to incorporate vocal tract lengths suitable for children by performing frequency warping. 25 warping factors between 1.0 and 2.0, the latter corresponding to halving the vocal tract length, were used, resulting in 25 instances of each training utterance. A multi-warp model was trained on all the warped data.

Warp factor specific phone models were also trained on the corresponding subset of the warped training speech. This set of models was used both to perform utterance-based warping comparable with standard VTLN and phoneme-dependent warping. In the second case, the search procedure consisted of determining a warp factor for each phoneme separately while all other factors were set to a fixed default value.

The number of mixture components in each state was empirically optimized to 32 using a development set of 60 adult speakers. The same number was used in the multi-warp model in order to keep its complexity and recognition time equal to those of the adult model. A word insertion penalty was also optimized on the development set.

# Results

A comparison of recognition results of an adult model, a number of vocal tract compensation methods on the adult model, and a child model is shown in Figure 1. Extending the training data by applying knowledge of vocal tract length implications decreased the number of errors by 50%. All vocal tract length methods performed approximately equally well in terms of word error rate. However, the multi-profile model trained on all the warped training data was superior in terms of execution time. This method needed a single invocation of the recognition procedure instead of 25 and 500 for the phone-independent and dependent methods respectively. The reason is that an explicit search for the best warping factor is not needed with this model.

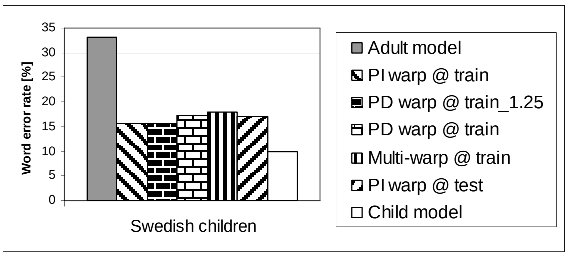

*Figure 1. Results for different phone-independent (PI) and -dependent (PD) vocal tract compensation methods. PI warping was performed on the training or test data. Two starting factor values for warp factor search was evaluated for PD. The Multi-warp bar represents the case when all warp factors are applied during training and no search during testing. Adult and child trained models with no compensation are used as baselines.*

A more detailed analysis of the results is shown in Figure 2. Traditional VTLN on the test utterance and utterance-based VTLA follow each other closely with some indication of less errors of VTLA for the youngest children.

Phone-independent warping resulted in slightly less errors than phone-dependent warping. This is in line with Potamianos' and Narayanan's prediction (2003) that warp estimation would be an issue in phone-dependent warping. The different results when two different starting points were chosen for phoneme-dependent warping illustrate that the implemented search algorithm is not guaranteed to find a global maximum point.

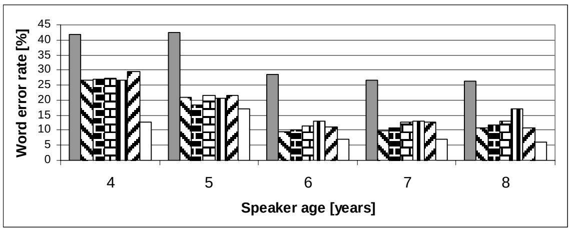

*Figure 2. Recognition results for different age groups. The methods and legends are identical to those in Figure 1.*

# Discussion and Conclusion

The proposed phoneme-dependent warping procedure performed at similar levels as the conventional phoneme-independent VTLN technique. Also the multi-warp model performed similarly well as standard VTLN. The latter result is in line with findings of mixing adult and children's speech (Gerosa, Giuliani and Brugnara, 2005; Elenius and Blomberg 2005) to also include that of child speech prediction using frequency warping. The multi-warp technique lowers the computational time to a fraction of traditional VTLN.

The results show that data-driven techniques and existing acoustic-phonetic knowledge can be combined to produce acoustic models which extend the range of variability of the used training data. The continued project work will improve the search algorithms and study other knowledge sources for implementation.

# Acknowledgement

This work was partly financed by the Swedish Research Council.

# References

Batliner A, Blomberg M, D'Acry S, Elenius D and Giuliani D (2005). The PF_STAR Children's Speech Corpus. *InterSpeech* 2005, 2761 - 2764.

Elenius D and Blomberg M (2005). Adaptation and Normalization Experiments in Speech Recognition for 4 to 8 Year old Children. In Proc *Interspeech* 2005, 2749 - 2752.

Gerosa M, Giuliani D and Brugnara F (2005). Speaker Adaptive Acoustic Modeling with Mixture of Adult and Children's Speech. *Interpseech 2005*. 2193 – 2196.

Großkopf B, Marasek K, v. d. Heuvel, H., Diehl F, and Kiessling A (2002). SpeeCon - speech data for consumer devices: Database specification and validation. *Second International Conference on Language Resources and Evaluation 2002.*

Lee L and Rose R (1996). Speaker Normalization Using Efficient Frequency Warping Procedures. *Proc. Int. Conf. on Acoustic, Speech and Signal Processing*, 1996, 353-356.

Lindberg B, Johansen F T, Warakagoda N, Lehtinen G, Kacic Z, Zgank A, Elenius K and Salvi G (2000). A noise robust multilingual reference recogniser based on SpeechDat(II), *ICSLP 2000*. 370-373

Moore R (2003). A Comparison of the Requirements of Automatic Speech Recognition Systems and Human Listeners, *Eurospeech 2003*, 2581-2584.

Potamianos A and Narayanan S (2003). Robust Recognition of Children's Speech. *IEEE Transactions on Speech and Audio Processing*, 2003, 603 - 616

Welling L, Kanthak S and Ney H (1999). Improved Methods for Vocal Tract Normalization. *ICASSP 99*, 161-164.

Young S, Evermann G, Gales M, Hain T, Kershaw D, Moore G, Odell J, Ollason D, Povey D, Valtchev V and Woodland P (2005). The HTK book. *Cambridge University Engineering Department 2005*.