Generating Synthetic Children's Acoustic Models from Adult Models

Andreas Hagen, Bryan Pellom, and Kadri Hacioglu

Rosetta Stone Labs
{ahagen, bpellom, khacioglu}@rosettastone.com

Abstract

This work focuses on generating children's HMM-based acoustic models for speech recognition from adult acoustic models. Collecting children's speech data is more costly compared to adult's speech. The patent-pending method developed in this work requires only adult data to estimate synthetic children's acoustic models in any language and works as follows: For a new language where only adult data is available, an adult male and an adult female model is trained. A linear transformation from each male HMM mean vector to its closest female mean vector is estimated. This transform is then scaled to a certain power and applied to the female model to obtain a synthetic children's model. In a pronunciation verification task the method yields 19% and 3.7% relative improvement on native English and Spanish children's data, respectively, compared to the best adult model. For Spanish data, the new model outperforms the available real children's data based model by 13% relative.

1 Introduction

Language learning is becoming more and more important in the age of globalization. Depending on their work or cultural situation some people are confronted with various different languages on a daily basis. While it is very desirable to learn languages at any age, language learning, among other learning experiences, is comparably simpler for children than for adults and should therefore be encouraged at early ages.

Even though the children's language learning market is highly important, comprising effective speech recognition tools for pronunciation assessment is relatively hard due to the special characteristics of children's speech and the limited availability of children's speech data in many languages in the speech research community. Adult speech data is usually easier to obtain. By understanding the characteristics of children's speech the unconditional need for children's speech data can be lessened by altering adult acoustic models such that they are suitable for children's speech.

Children's speech has higher pitch and formants than female speech. Further, female speech has higher pitch and formants than male speech. Children's speech is more variable than female speech, and, as research has shown, female speech is more variable than male speech (Lee et al., 1999). Given this transitive chain of argumentation, the transformation from a male to a female acoustic model can be estimated for a language and applied (at a certain adjustable degree) to the female model. This process results in a synthetic children's speech model designed on the basis of the female model. Therefore, for a new language an effective synthetic children's acoustic model can be derived without the need of children's data (Hagen et al., 2008).

2 Related Work

Extensive research has been done in the field of children's speech analysis and recognition in the past few years. A detailed overview of children's speech characteristics can be found in (Lee et al., 1999). The paper presents research results showing the higher variability in speech characteristics among children compared to adult speech. The properties of children's speech that were researched were duration of vowels and sentences, pitch, and formant locations.

When designing acoustic models specially suited for children, properties as the formant locations and higher variability of children's speech need to be accounted for. The best solution for building children's speech models is to collect children's speech data and to train models from scratch (Ha-

Proceedings of NAACL HLT 2009: Short Papers, pages 77–80,
Boulder, Colorado, June 2009.

gen et al., 2003, Cosi et al. 2005). Researchers have also tried to apply adult acoustic models using speaker normalization techniques to recognize children's speech (Elenius et al., 2005, Potamianos et al. 1997). Adult acoustic models were adapted towards children's speech. A limited amount of children's speech data was available for adaptation. In (Gustafson et al., 2002) children's voices were transformed before being sent to the recognizer using adult acoustic models. In (Claes et al., 1997) children's acoustic models were built based on a VTL adaptation of cepstral parameters based on the third formant frequency. The method showed to be effective for building children's speech models.

3 Building Synthetic Children's Models from Adult Models

As mentioned in Section 1, research has shown that pitch and formants of children's speech are higher than for female speech. Female speech has higher pitch and formants than male speech. In order to exploit these research results a transformation from a male acoustic model to a female acoustic model can be derived. This transformation will map a male model as close as possible to a female model. The transformation can be adjusted and applied to the female model. The resulting synthetic model can be tested on children's data.

Parameters that are subject to transformation in this process are the mean vectors of the HMM states. The transformation can be represented as a square matrix in the dimension of the mean vectors. The transformation chosen in this approach is therefore linear and is for example capable of representing a vocal tract length adaptation as it was shown in (Pitz et al., 2005). Linear transformations (i.e. matrices) are also chosen in adaptation approaches as MAPLR and MLLR, whose benefit has been shown to be additive to the benefit of VTLN in speaker adaptation applications. A linear transform in the form of a matrix is therefore well suited due to its expressive power as well as its mathematical manageability.

3.1 Transformation Matrix

The transformation matrix used in this approach is estimated by mapping the male to the female acoustic model, such that each HMM state mean vector in the male model is assigned a corresponding mean vector in the female model. Information used in the mapping process is the basic phoneme and context. The resulting mean vector pairs are used as source and target features in the training process of the transformation matrix. During training the matrix is initialized as the identity matrix and the estimate of the mapping is refined by gradient descent. In a typical acoustic model there are several hundred, sometimes thousands, of these mean vector pairs to train the transformation matrix. The expression that needs to be minimized is:

$$T = \arg\min_{A} \sum_{(x,y)\,pairs} (Ax - y)^2$$

where T is the error-minimizing transformation matrix; *x* is a male model's source vector and *y* it corresponding female model's target vector.

In this optimization process the Matrix *A* is initialized as the identity matrix. Each matrix entry a_{ij} is updated (to the new value $a_{ij}^{'}$) in the following way by gradient descent:

$$a_{ij}^{'} = a_{ij} + k\left(A_i x - y_i\right)x_j$$

where A_i is the *i*-th line of matrix *A* and *k* determines the descent step size (*k<0* and incorporates the factor of 2 resulting from the differentiation). The gradient descent needs to be run multiple times over all vector pairs *(x,y)* for the matrix to converge to an acceptable approximation which is called the transformation matrix *T*.

3.2 Synthetic Children's Model Creation

The transformation matrix can be applied to the female model in order to create a new synthetic acoustic model which should suit children's speech better than adult acoustic models. It is unlikely that the transformation applied "as is" will result in the best model possible, therefore the transformation can be altered (amplified or weakened) in order to yield the best results. An intuitive way to alter the impact of the transformation is taking the matrix *T* to a certain power *p*. Synthetic models can be created by applying T^p to the female model[1], for various values *p*. If children's data is available for evaluation purposes, the best value of *p* can be determined. The power *p* is claimed to be language independent. It might vary in nuances, but experi-

[1] Taking a matrix to the power of *p* is meant in the sense $T^{p^{1/p}} = T$, $T^0 = Identity$, $T^1 = T$

ments have shown that a value around 0.25 is a reasonable choice.

3.3 Transformation Algorithm

The previous section presented the theoretical means necessary for the synthetic children's model creation process. The precise, patent-pending algorithm to create a synthetic children's model in a new language is as follows (Hagen et al., 2008):

1. Train a male and a female acoustic model
2. Estimate the transform T from the male to the female model
3. Determine the power p by which the transform T should be adjusted
4. Apply T^p to the female acoustic model to create the synthetic children's model

Step 3, the determination of the power p, can be done in two different ways. If children's test data in the relevant language is available, various models based on different p-values can be evaluated and the best one chosen. If there is no children's data available in a new language, p can be estimated by evaluations in a language where there is enough male, female, and children's speech data available. The claim here is that the power p is relatively language independent and estimating p in a different language is superior to a simple guess.

4 Experiments

The algorithm was tested on two languages: US English and Spanish. For both languages sufficient male, female, and children's speech data was available (more than 20 hours) in order to train valid acoustic models and to have reference children's acoustic models available. For English test data we used a corpus of 22 native speakers in the age range of 5 to 14. The number of utterances is 2,182. For Spanish test data the corpus is comprised of 19 speakers in the age range of 8 to 13 years. The number of utterances is 2,598.

The transform from the male to the female model was estimated in English. The power of p was gradually increased and the transformation matrix was adjusted. With this adjusted matrix T^p a synthetic children's model was built. This synthetic children's model was evaluated on children's test data and the results were compared to the reference children's model's and the female model's performance.

When speech is evaluated in a language learning system, the first step is utterance verification, meaning the task of evaluating if the user actually tried to produce the desired utterance. The Equal Error Rate (EER) on the utterance level is a means of evaluating this performance. For each utterance an in- and out-of-grammar likelihood score is determined. The EER operating points, determined by the cutting point of the two distributions (in-grammar and out-of-grammar), are reported as an error metric. Figure 1 shows the EER values of the synthetic model applied to children's data.

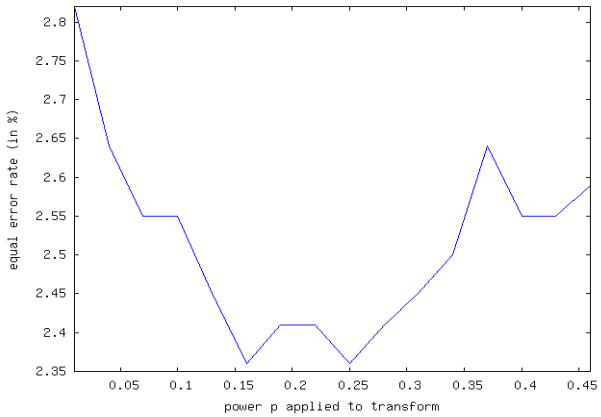

Figure 1: Synthetic model's EER performance depending on the power p used for model creation.

It can be seen that the best performance is reached at about $p=0.25$. The overview of the results is given in Table 1.

	Equal Error Rate
Real Children's Model	1.90%
Male Model	4.07%
Female Model	2.92%
Synthetic Model	2.36%

Table 1: EER numbers when using a real children's model compared to a male, female, and synthetic model for children's data evaluation.

The results show that the synthetic children's model yields good classification results when applied to children's data. The gold standard, the real children's model application, results in the best EER performance.

If the same evaluation scenario is applied to Spanish, a very similar picture evolves. Figure 2 shows the EER results versus transformation power p for Spanish children's data.

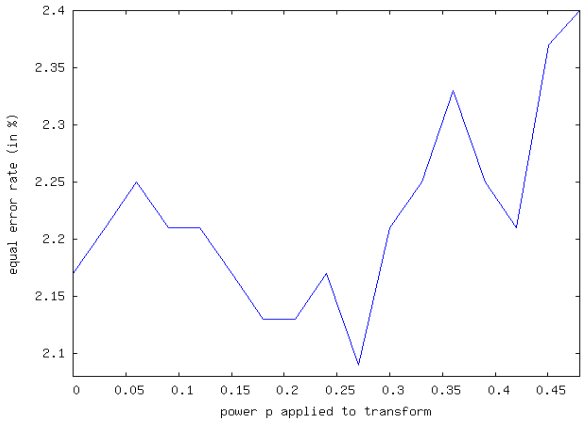

Figure 2: Spanish synthetic model's EER performance depending on the power p used for model creation.

In Figure 2 it can be seen that the optimal setting for *p* is about *0.27*. This value is very similar to the one found for US English, which supports, but certainly does not prove, the language independence claim. Results for Spanish are given in Table 2.

	Equal Error Rate
Real Children's model	2.40%
Male model	5.62%
Female model	2.17%
Synthetic model	2.09%

Table 2: EER numbers for Spanish when using a real children's model compared to a male, female, and synthetic model for Spanish children's data evaluation.

Similar to English, the Spanish synthetic model performs better than the female model on children's speech. Interestingly, the acoustic model purely trained on children's data performs worse than the female and the synthetic model. It is not clear why the children's model does not outperform the female and the synthetic model; an explanation could be diverse and variable training data that hurts classification performance.

It can be seen that for US English and Spanish the power *p* used to adjust the transformation is about *0.25*. Therefore, for a new language where only adult data is available, the transformation from the male to the female model can be estimated and applied to the female model (after being adjusted by *p=0.25*). The resulting synthetic model will work reasonably well and could be refined as soon as children's data becomes available.

5 Conclusion

This work presented a new technique to create children's acoustic models from adult acoustic models without the need for children's training data when applied to a new language. While it can be assumed that the availability of children's data would improve the resulting acoustic models, the approach is effective if children's data is not available. It will be interesting to see how performance of this technique compares to adapting adult models by adaptation techniques, i.e. MLLR, when limited amounts of children's data are available. Two scenarios are possible: With increasing amount of children's data speaker adaptation will draw even and/or be superior. The other possibility is that the presented technique yields better results regardless how much real children's data is available, due to the higher variability and noise-pollution of children's data.

References

Claes, T., Dologlou, I, ten Bosch, L., Van Compernolle, D. 1997. *New Transformations of Cepstral Parameters for Automatic Vocal Tract Length Normalization in Speech Recognition*, 5th Europ. Conf. on Speech Comm. and Technology, Vol. 3: 1363-1366.

Cosi, P., Pellom, B. 2005. *Italian children's speech recognition for advanced interactive literacy tutors*. Proceedings Interspeech, Lisbon, Portugal.

Elenius, D. and Blomberg, M. 2005. *Adaptation and Normalization Experiments in Speech Recognition for 4 to 8 Year old Children*. Proceedings Interspeech, Lisbon, Portugal.

Gustafson, J., Sjölander, K. 2002. *Voice transformations for improving children²s speech recognition in a publicly available dialogue system*. ICSLP, Denver.

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

Lee, S., Potamianos, A., and Narayanan, S. 1999. *Acoustics of children's speech: Developmental changes of temporal and spectral parameter*. J. Acoust. Soc. Am., Vol. 105(3):1455-1468.

Pitz, M., Ney, H. 2005. *Vocal Tract Normalization Equals Linear Transformation in Cepstral Space.* IEEE Trans. Speech & Audio Proc., 13(5): 930-944.

Potamianos, A., Narayanan, S., and Lee, S. 1997. *Automatic Speech Recognition for Children*. Proceedings Eurospeech, Rhodes, Greece.

Hagen, A., Pellom, B., and Hacioglu, K. 2008. *Method for Creating a Speech Model*. US Patent Pending.