Acoustic Normalization of Children's Speech

Georg Stemmer, Christian Hacker, Stefan Steidl, Elmar Nöth

Universität Erlangen–Nürnberg
Lehrstuhl für Mustererkennung (Informatik 5)
Martensstr. 3, 91058 Erlangen, Germany
stemmer@informatik.uni-erlangen.de

Abstract

Young speakers are not represented adequately in current speech recognizers. In this paper we focus on the problem to adapt the acoustic frontend of a speech recognizer which has been trained on adults' speech to achieve a better performance on speech from children. We introduce and evaluate a method to perform non-linear VTLN by an unconstrained data-driven optimization of the filterbank. A second approach normalizes the speaking rate of the young speakers with the PSOLA algorithm. Significant reductions in word error rate have been achieved.

1. Introduction

1.1. Motivation

Very young speakers are not represented adequately in current speech recognizers. This is partly due to unbalanced amounts of training data and partly caused by the acoustic features and models. The large acoustic variability of children's speech together with a change of the location and range of the optimal parameter values need specialized feature extraction methods and acoustic models.

The influence of the speaker's age on the accuracy of a speech recognizer has firstly been investigated by J. Wilpon and C. Jacobsen in [1]. The error rate of a speech recognizer which has been trained with data from speakers of all ages increases significantly for speakers which are twelve years old or younger. It is also shown that best performance can be achieved when each age group is represented by an adequate amount of data in the training set, however, the recognition is still worse for elderly speakers and children. In [2] a detailed analysis of the relation between the word error rate of a speech recognizer and the age of young speakers shows that the recognition performance for children's speech is up to four times worse than for adults and that adult levels are reached around thirteen or fourteen years of age. As a main portion of children's speech is located in relatively high-frequency regions of the spectrum it has been observed that recognition quality is degraded much more than speech from adults by the effects of bandwidth reduction, e.g. for telephone speech [3].

1.2. Acoustic characteristics of children's speech

S. Lee *et al.* analyze in [4] duration, pitch, formant frequencies and the spectral envelope of children's speech. It is found that durations of certain vowels are longer for young children than for adults. The children have a higher variability of duration between the vowels, which stabilizes around eleven. Up to age twelve there is no difference in the development of a speaker's pitch between males and females. There is a continuous decrease in pitch between seven and twelve. Male speakers have a steep pitch drop in puberty between twelve and 15. The separation between male and female formant frequencies begins around ten and is finished around 15. There is a linear relationship between the speaker's age and the vowel-dependent formant frequencies for male but not for female speakers. Similar to the observations for the vowel durations, young children do also have a higher spectral variation and formant variability than adults. This may be due to the fact that children are less-skilled in coarticulation [4].

1.3. Approach

In this paper we focus on the problem to adapt the acoustic frontend of a speech recognizer which has been trained on adults' speech to achieve a better performance when speech from children has to be recognized. We will investigate if our approach can also lead to improvements when we use children's speech to train the recognizer.

Two aspects of children's speech are investigated: Firstly, we normalize the spectral characteristics of children's speech by a generalization of *Vocal-Tract Length Normalization (VTLN)*. VTLN is a linear or bilinear scaling of the frequency axis [5, 6, 7]. Based on a non-linear extension of VTLN we try to find out if an optimal filter bank for extraction of acoustic features from children's speech can be derived directly from the data with as little constraints as possible. Secondly we evaluate if the recognition rate can be improved by normalization of the children's speaking rate. For both approaches word error rates on a corpus of read speech from children are given. Finally, we discuss the results and give an outlook on our work in the near future.

1.4. Related work

Several approaches to improve the acoustic frontend for the automatic recognition of children's speech can be found in the literature; most of them are based on some kind of VTLN. A. Potamianos and R. Rose show in [8] a strong relationship between the optimal warping factor and the age of the speaker. VTLN provides relative reductions in word error rate of up to 60% when a speech recognizer which has been trained on male adult speakers is tested on children's speech [8]. D. Burnett and M. Fanty report similar experiments in [9], however the frequency axis is not scaled but *shifted*. Experiments by Potami-

Parts of this work were funded by the German Federal Ministry of Education, Science, Research and Technology (*BMBF*) in the framework of the SmartKom project under Grant 01 IL 905 K7 and by the European Commission (IST programme) in the framework of the PF-STAR project under Grant IST-2001-37599. The responsibility for the content lies with the authors.

anos *et al.* on phoneme-dependent warping factors in [2] indicate that the phoneme-independent factor is a valid approximation as the influence of the age on the parameter is similar between all phonemes. Additional improvements could be gained by extending VTLN by using two warping factors instead of only one [2]. An interesting alternative has been developed by J. Gustafson and K. Sjölander [10]: A voice transformation method is applied directly to the speech signal in order to compress the spectrum and to change the speaker's pitch. The compression is equivalent to the scaling of the frequency axis during feature computation but can be integrated into any speech processing system even if the speech recognizer itself cannot be accessed for technical reasons.

2. Non-linear VTLN

VTLN is based on the assumption that a significant part of the inter-speaker variation in the features can be eliminated by a speaker-dependent linear scaling of the frequency axis. It may not be optimal to constrain the scaling to be linear. As most applications need a very rapid adaption to the current speaker it would be too costly in terms of computational effort and adaption data to perform an extensive search for the optimal non-linear frequency scaling. In this paper, however, the frequency scaling is determined at once for a whole speaker group as we are more interested in the acoustic differences between the voices of children and adults than in the individual differences between speakers. This means that the frequency scaling can be computed offline and does not need to be performed in real time. In the following we introduce a data-driven algorithm to determine an arbitrary non-linear scaling of the frequency axis, which we want to call *non-linear VTLN*. The scaling of the frequency axis is performed by optimizing the Mel-curve which determines the filterbank for the computation of the Mel-frequency cepstral coefficients (MFCC).

For this purpose we have to define a suitable parameterized representation of the Mel-curve. The Mel-curve is given by

$$mel(f) = 1125 \cdot \ln(1 + \frac{f}{700}) \qquad (1)$$

In order to adjust the curve to the properties of the data we determine n sampling points $(f_i, p(f_i))$ which are uniformly spaced in the Mel-scale:

$$f_i = mel^{-1}((i-1) \cdot \frac{f_{max}}{n-1}), \;\; i \in 1, \ldots, n \qquad (2)$$

f_{max} depends on the sampling frequency and is in our case 8000 Hz. The initialization of the sampling points is given by the Mel-curve itself:

$$p(f_i) = mel(f_i), \;\; i \in 1, \ldots, n \qquad (3)$$

The optimized curve $opt(f)$ is a cubic spline curve which interpolates the sampling points $(f_i, p(f_i))$. The optimization algorithm iteratively changes the values $p(f_i)$ in order to meet an optimization criterion. We decided to employ the simplex algorithm [11] which does not need the computation of a gradient. The optimization criterion of the simplex algorithm is to maximize the recognition rate of a Gaussian classifier for sub-phonetic labels. Preliminary experiments had shown a good correlation between the recognition rate of the classifier and the word error rate of a speech recognizer which has been trained on the same data.

The following paragraph summarizes the algorithm for the data-driven optimization of the Mel-curve:

1. initialize f_i and $p(f_i)$ according to Eq. 2 and Eq. 3
2. iteratively adjust $p(f_i)$ using the simplex algorithm and the following optimization criterion:
   (a) compute interpolating cubic spline curve $opt(f)$ for points $(f_i, p(f_i))$
   (b) determine filterbank according to $opt(f)$
   (c) extract acoustic features using the new filterbank
   (d) train Gaussian classifier for sub-phonetic labels
   (e) evaluate Gaussian classifier, get recognition rate
   (f) pass recognition rate to simplex algorithm
3. return the optimal $opt(f)$

Note that the use of an interpolating spline function in contrast to directly optimizing the filter bank parameters has the advantage that the number n of parameters which have to be optimized can be chosen independently from the number of the filters which are used in the feature extraction. A major disadvantage of the proposed algorithm is the computational effort: Each iteration of the simplex algorithm evaluates the optimization criterion several times. For our data set this resulted in 3-4 hours of computation time on a Pentium 4 computer until the local optimum was reached.

3. Normalization of the speaking rate

As the duration of certain vowels in children's speech is longer than for adults, we considered a normalization of the speaking rate. Our approach is motivated by the work from J. Gustafson and K. Sjölander [10] who applied the *Pitch-Synchronous Overlap and Add (PSOLA)* algorithm to normalize the pitch of children's speech. The PSOLA algorithm is widely used in speech synthesis for the manipulation of pitch and duration of a speech signal. Preliminary experiments had shown that in our case manipulation of the pitch did not yield additional improvements. Note that in [10] effects of pitch manipulation and spectral compression are not evaluated separately. We apply the implementation of PSOLA which is part of the PRAAT software (http://www.praat.org) to change the duration of the children's utterances while keeping the original pitch.

4. Data

The children's speech corpus consists of 3.5 h read speech data of 62 children (29 male and 33 female) at the age of 10 to 12 years. The pupils read four different German texts: *Nordwind und Sonne (The North Wind and the Sun)* and three texts of the reading test *Zürcher Lesetest* [12]. Each text is about 90 words long. In order to judge the reading capabilities of each child we asked two master-level students of psychology to rate each pupil's reading for each of the three *Zürcher Lesetest* texts. Both a fluency rating and a rating of the reading expression are on a scale from 1 (best) to 5 (worst). No pupil got the rating 5 for any of the texts. The ratings were averaged over all three texts which gives a rating between 1-4 for each child in the categories fluency and reading expression.

Tab. 1 shows how the data has been partitioned for the speech recognition experiments. As we wanted speakers and read texts to be disjunct between training and evaluation set, the *Nordwind und Sonne* readings of the training speakers and the *Zürcher Lesetest* readings of the evaluation speakers are not used for any of the experiments.

task	sentences	speakers	read text
training	920	40	*Zürcher Lesetest*
validation	46	2	*Zürcher Lesetest*
evaluation	120	20	*Nordwind und Sonne*

Table 1: *Partitioning of the children's speech corpus for the training and evaluation of the speech recognizer.*

For the training of the adults' speech recognizer a subset of the speech database from the VERBMOBIL project [13] was used. It consists of 28 hours of spontaneous dialogues between adult speakers in German (11762 turns of 610 dialogues).

5. Baseline systems

Two different baseline systems are available: An adults' speech recognizer and a recognizer for children's speech. The acoustic models of the adults' speech recognizer have been trained on the data from the VERBMOBIL project while the acoustic models of the children's speech recognizer have been estimated on the training set of the children's speech corpus as shown in Tab. 1. Both recognizers have the same recognition vocabulary of 71 words which are the words from the *Nordwind und Sonne* text. The children sometimes gave comments to the text or made reading errors so 2.1% of the spoken words are not contained in the vocabulary. As our main interest is to measure the performance of the acoustic frontend both recognizers use only an unigram language model which is estimated on the written *Nordwind und Sonne* text (not on the spoken word sequence). The adults' speech recognizer has a word error rate (WER) of 32.5% on the evaluation set. The children's speech recognizer achieves a WER of 18.5% on the same data.

6. Experiments and results

6.1. Non-linear VTLN

The goal of the feature normalization experiments is to find a feature representation which gives the best recognition performance of the adults' speech recognizer on children's speech data. The recognizer itself is not altered. In order to represent this setting in the optimization criterion, the Gaussian classifier is trained only once on the VERBMOBIL dataset with the unmodified MFCC features. For the computation of the optimization criterion the recognition rate of the Gaussian classifier is evaluated on the training and the validation subsets of the children's speech corpus based on features which have been extracted with the optimized Mel-curve. We set $n = 12$; the first and the last sampling point $(f_1, p(f_1))$ and $(f_n, p(f_n))$ are fixed which gives 10 free parameters for the simplex optimization. The simplex algorithm needs 494 evaluations of the optimization criterion. The optimized Mel-curve is shown in Fig. 1. The adults' speech recognizer achieves a WER of 25.5% on the evaluation set with the improved feature extraction. This corresponds to a relative reduction of more than 20%. When only the two speakers from the validation set are used together with 8 free parameters for the non-linear VTLN optimization the corresponding WER is already 25.9%. It can be seen clearly in Fig. 1 that the optimized Mel-curve is the result of a non-linear transformation of the frequency scale. However, this is not a proof that the optimal VTLN is non-linear as the curve in Fig. 1 represents just a local optimum and there may be linear scalings which perform better. We therefore optimized the linear VTLN on the *evaluation set* by varying the warping factor between 0.8 and 1.48 in steps of 0.02. The best WER that can be achieved by the linear VTLN on the test set is 25.6%. This is still slightly worse than the non-linear VTLN, i.e. no linear frequency scaling can do better than the non-linear VTLN on the evaluation set. As the difference is not significant we cannot claim that this is a proof that the non-linear VTLN is better than its linear counterpart, however we believe that it is a strong indication of this statement.

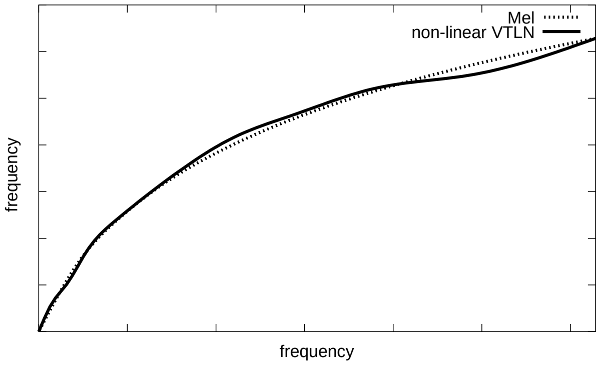

Figure 1: *Standard Mel-curve and improved Mel-curve. The new Mel-curve has been optimized with non-linear VTLN using the adults' speech recognizer.*

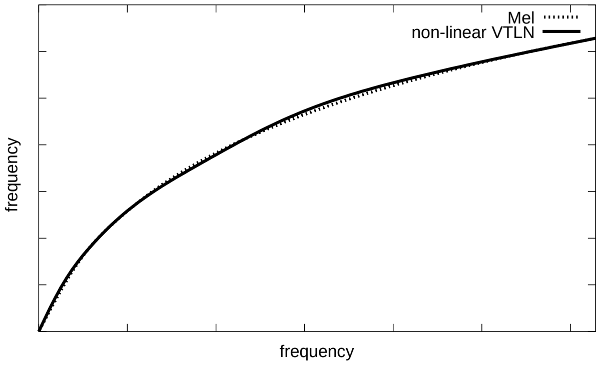

Figure 2: *Standard Mel-curve and improved Mel-curve. The new Mel-curve has been optimized with non-linear VTLN using the children's speech recognizer.*

We also evaluated if the non-linear VTLN can be used to find a better Mel-curve for the children's speech recognizer. The Gaussian classifier is estimated on the training set of the children's speech corpus and evaluated on the two speakers of the validation set. Both for the training and the evaluation the feature extraction is based on the modified Mel-curve. The number of sampling points is set to 10 which makes 8 free parameters in the simplex optimization. The resulting optimized Mel-curve after 93 evaluations of the optimization criterion is shown in Fig. 2. It can be seen easily that the differences to the standard Mel-curve are only very small. This is no surprise as the Gaussian classifier has been evaluated on the data of only two speakers. The corresponding WER on the evaluation set is 17.3%, i.e. as one may expect from Fig. 2 there is no significant improvement over the baseline.

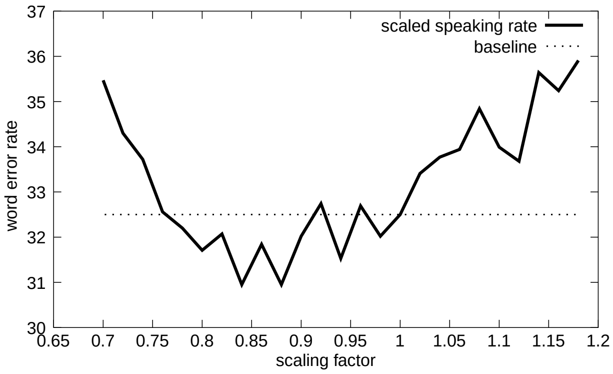

Figure 3: *Relationship between the scaling of the speaking rate and the WER of the adults' speech recognizer on the evaluation data.*

6.2. Normalization of the speaking rate

In the following we measure the maximum possible WER reduction that can be achieved by normalization of the speaking rate. As we currently have no criterion to determine the optimal scaling factor for a speaker all experiments are performed on the evaluation set only. In a first experiment the same scaling factor is applied to all speakers. Fig. 3 shows the WER on the evaluation set when the scaling factor is varied in the range between 0.7 and 1.18 in steps of 0.02. Small factors accelerate the speaking rate in the transformed speech signal; large factors decrease the speaking rate. The minimum of the curve is clearly below 1.0, i.e. acceleration of the children's speaking rate reduces the WER. The best achievable WER in this case is 30.9% (for scaling factors 0.84 and 0.88). We noticed a very large speaker dependency of the optimal scaling factor. When the optimal scaling factor is chosen individually for each speaker a WER of 28.6% can be reached. One piece of information that could be used to determine the optimal scaling factor for a speaker in advance is his or her reading capability. It can be expected that bad readers speak more slowly than good readers. The correlation between optimal scaling factor of a speaker and the corresponding expert fluency rating (1=best – 5=worst) which has been estimated on the training set is -0.40, i.e. speakers who get good ratings have scaling factors closer to 1.0. The same holds for the correlation between optimal scaling factor of a speaker and the rating of the reading expression (1=best – 5=worst) which is -0.41.

7. Conclusion and future work

We introduced and evaluated two new ways to increase speech recognition performance for children: Non-linear VTLN and normalization of the speaking rate with the PSOLA algorithm. Both methods provide reductions in word error rate for a speech recognizer which has been trained on adults' speech. However, we were not able to find a better "Mel-scale for children" which does also improve performance when the recognizer is already trained on speech from children. One reason for this is surely a lack of data.

In the near future we plan to repeat the experiments on a larger database. For this purpose we have recorded read speech from about 50 German pupils in the age between 10-12, each of them reading 15 min German and 5 min English texts. Additional recordings of spontaneous and emotional speech have been made in Wizard-of-Oz experiments: children talked to Sony's entertainment robot AIBO and instructed it to fulfill selected tasks.

In further experiments we will look for a suitable criterion which can be used to determine the optimal scaling factor for the speaking rate of a speaker. Additional improvements can be expected when non-linear VTLN is combined with the speaking rate normalization. We also think about applying the algorithm which we developed for the non-linear VTLN to optimize the $\log$-scale in the MFCC computation.

8. References

[1] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1996, pp. 349–352.

[2] A. Potamianos, S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Proc. European Conference On Speech Communication and Technology (EUROSPEECH)*, vol. 5, 1997, pp. 2371–2374.

[3] Q. Li and M. Russell, "Why is automatic speech recognition of children's speech difficult?" in *Proc. European Conference On Speech Communication and Technology (EUROSPEECH)*, 2001, pp. 2671–2674.

[4] S. Lee, A. Potamianos, and S. Narayanan, "Analysis of children's speech: Duration, pitch and formants," in *Proc. European Conference On Speech Communication and Technology (EUROSPEECH)*, 1997, pp. 473–476.

[5] A. Andreou, T. Kamm, and J. Cohen, "Experiments in vocal tract length normalization," in *Proceedings of the CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[6] L. Lee and R. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1996, pp. 353–356.

[7] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, 1996, pp. 346–348.

[8] A. Potamianos and R. Rose, "On combining frequency warping and spectral shaping in HMM based speech recognition," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1997.

[9] D. Burnett and M. Fanty, "Rapid unsupervised adaption to children's speech on a connected-digit task," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 1996, pp. 1145–1148.

[10] J. Gustafson and K. Sjölander, "Voice transformations for improving children's speech recognition in a publicly available dialogue system," in *Proc. International Conference On Spoken Language Processing (ICSLP)*, 2002, pp. 297–300.

[11] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C.* Cambridge: Cambridge University Press, 1992.

[12] M. Linder and H. Grissemann, *Zürcher Lesetest*, 6th ed. Testzentrale Göttingen, 2000.

[13] W. Wahlster, *Verbmobil: Foundations of Speech-to-Speech Translation.* New York, Berlin: Springer, 2000.