Spectral and TRAP-Based Characterization of Children's Speech

Christian Hacker, Stefan Steidl, Elmar Nöth, Anton Batliner

Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Germany
hacker@informatik.uni-erlangen.de

Abstract

Recognition of young speakers causes problems in automatic systems. In this paper spectral and temporal, TRAP based, characteristics of adults' and children's speech will be analyzed. Two German databases with spontaneous speech of both speaker groups and an American read children's speech corpus are compared. Speaker variability and recognition performance per frequency band are investigated. In phone recognition, TRAP and MFCC based features prove to contain a high amount of complementary information; TRAPs perform even better on a small vocabulary corpus. For word recognition they could not yet outperform standard features computed on an optimal filterbank. However, decoded with small codebooks only, the results motivate further research.

1. Introduction

1.1. Motivation and approach

Most current speech recognizers are trained to reach good performance for adults. Other speaker groups like young, elderly or non-native speakers cannot be recognized robustly enough. In the following we will focus on children's speech. In different publications various techniques are introduced how to preprocess the signals in order to achieve better recognition results whenever speech is decoded with a recognizer trained on adults. One very common approach is vocal tract length normalization (VTLN) [1]. If enough prerecorded children's speech were available, another possibility would be to train a special recognizer for children. However, it is reported that even then speech of young speakers cannot be decoded as robustly as adults' speech [2]. Reasons are the higher spectral variability or higher variability in speaking rate, vocal effort and degree of spontaneity [1]. The varieties in the fundamental and formant frequencies are caused by the different lengths of the vocal tracts; furthermore, children are less skilled in coarticulation.

In this paper speech and phone recognizers trained separately on children's or adults' speech are compared. The investigations are restricted to children with age 6 – 13. Spontaneous children's speech and read children's speech from young speakers with American and German mother tongue are analyzed and compared with adults' speech. Both spectral and temporal characteristics as well as their influence to phone classification and speech recognition are pointed out.

1.2. Related work

An overview of the characteristics of children's speech for different age groups and in comparison with adults' speech is given in Potamianos et al. [1]. Phone dependent intra-speaker variability which rises for younger children is analyzed with the cepstral distance measure. For frequency warping an algorithm to adapt the warping factor is introduced. VTLN of adults' and children's speech is applied in [3]. In [4, 5] an approach for nonlinear VTLN is developed and the speaking rate of read children's speech is analyzed. In [6] effects of wrong pronunciation are shown, formant frequencies and problems due to bandwidth reduction in telephone speech are analyzed.

The approach to recognize speech with temporal patterns is motivated and explained by Hermansky et al. in [7]. A TRAP is a vector representing the temporal evolution of phones in a critical band. 1 sec. time trajectories centered around the frame under consideration are taken into account. The mean TRAP is obtained by averaging all TRAPs belonging to the same phone regardless the context. For classification in a first step the TRAP of the current frame is compared with all mean TRAPs. In this way scores for all classes are obtained. Better scores are achieved by neural networks (NN). In a second step scores of 15 critical bands are combined by a huge NN. A combination of the output of the baseline system with the output of the TRAP based classifier improves recognition rates. Good performance is further obtained for noisy speech. In [8] TRAPs from adjacent spectral bands are combined.

2. Corpora

In this paper, three diverse corpora are explored. All data is sampled with 16 kHz. For the *Aibo* database (spontaneous German speech) we recorded 51 children which were playing with SONY's Aibo entertainment robot (age 10 – 13, 21 male, 30 female). The children gave spoken instructions to the Aibo in order to fulfill several tasks like to guide the Aibo around a map that was printed on a carpet. The children were told to talk to the robot like they would talk to a friend. They were led to believe that the Aibo was responding to his or her commands. However, it was actually being controlled by a human operator, using the "Aibo Navigator" software over a wireless LAN. This Wizard-of-Oz procedure was developed to elicit spontaneous and emotional children's speech. For details please refer to [9]. 9 hours of speech has been collected. The vocabulary contains 850 words (380 of them occur only once) and 350 word fragments. Thus, some phones that occur frequently can only be observed in a specific context. 8957 turns are used for training, 1381 for validation and 3453 for testing.

The *Youth*[1] database contains 14 hours of read American children's speech. Most of the 56 male and 79 female children (age 6 – 10) read about 200 of 406 phonetically rich phrases and single words. The total vocabulary is 780 words. We use 15180 turns for training, 4868 for validation and 4899 for evaluation.

To compare children's speech with adults' speech we employed a part of the German *Verbmobil* database (5 hours of spontaneous German speech, 49 male, 33 female). The training set consists of 950 turns, the validation set of 94 and the test set of 846 turns. This corpus was recorded for the Verbmobil

[1] Youth © 2002, Carnegie Speech Company, Inc.

project where bi-directional translation of spontaneous speech has been investigated. The vocabulary comprises 2675 words.

3. Recognition system

3.1. Short-time feature extraction

In this paper, we focus on the analysis of spectral and temporal feature extraction. For the short-time analyses the spectrum of 10 msec. time windows is calculated, smoothed with a Mel filterbank with triangular filters and afterwards logarithmized. The mean is calculated adaptively and subtracted. The Mel spectrum coefficients are decorrelated with the discrete cosine transform. Together with the short-time energy we yield 12 coefficients (MFCC) plus 12 derivatives (Δ) that are approximated by the regression within 5 short-time windows.

3.2. Labeling of frames

Based on these features we compute a baseline word recognizer for each database, that is employed for a forced frame-based alignment of the data. The labels are required for the computation of TRAPs as well as for the classification on the phone level (Sec. 3.3, 3.4). To obtain a more robust labeling only those frames are taken into account where two alignments of different recognizers based on different filterbanks agree. The labels are combined to 24 (German) resp. 23 (English) phone supersets to make it possible to compare phones in both languages.

3.3. Temporal patterns

For each of the Mel spectrum coefficients (Sec. 3.1) calculated with a filterbank with 18 filters, we consider in addition the temporal progression (TRAPS). Each TRAP is a vector of short-time filter energies in a time interval of about 1 sec (2×50 frames context). We reduced the dimension of the vector to 33 by taking not every component but all that are next to the TRAP's center and only some in the surrounding time context. With a Gaussian classifier for each TRAP scores for all phone supersets are computed. These TRAP specific score vectors are concatenated to a high dimensional vector and transformed into a subspace by the linear discriminant analysis (LDA). In this paper the result has dimension 12. We do not use NNs because the estimation of the LDA transformation is much less time consuming than a training procedure. For phone recognition the vector resulting from the LDA is input of a Gaussian classifier. For word recognition this vector and the one resulting from Sec. 3.1 are both applied to train HMMs.

3.4. Classification

For the phone classification we estimate one Gaussian density per class (phones or phone supersets) on the training data. The performance is measured by the recognition rate resp. by the mean score fired by the density of the reference phone. Both are weighted by the a-priori probability and averaged over all classes. The evaluation is done on the validation data set. For word recognition we use the ISADORA system [5] to train a polyphone based recognizer with semicontinuous HMMs. Word accuracy (WA) is evaluated on the test data set. In [10, 11, 5] an approach for the training of multiple codebooks for feature streams is described. Thereby the probabilities of the feature sub-vectors are exponentially weighted by stream dependent weights, which are fixed in all HMM states.

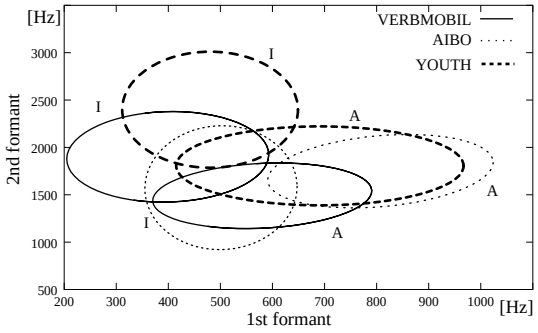

Figure 1: *1st and 2nd formant of the vowel supersets 'A' and 'I'*

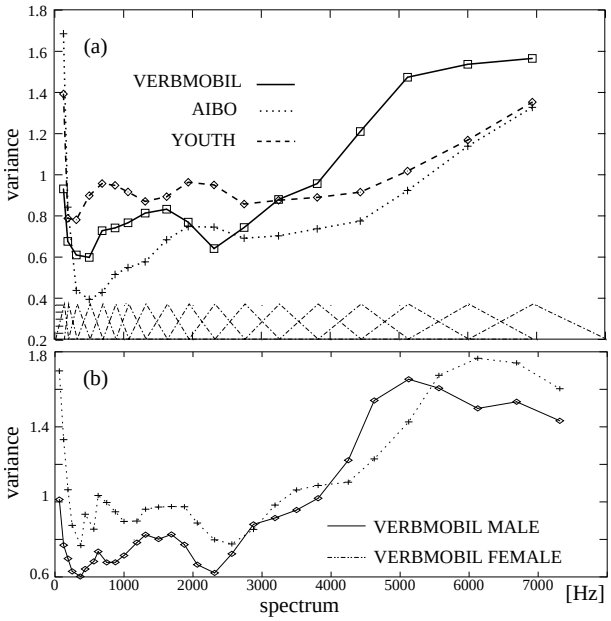

Figure 2: *Averaged speaker variability of the energy values in different Mel filterbanks (a) for the adults' and children's speech databases and (b) for male and female adult speakers.*

4. Experiments and results

4.1. Spectral and temporal characteristics

In some frequency bands consequences of higher variances of energy values may be better separability and higher score values obtained by a Gaussian classifier. In this section we have a look at characteristics and class-wise averaged scores obtained for phone supersets. First the spectral characteristics of the data will be analyzed. In Fig. 1 the distribution of the first and the second formants of the vowel supersets "A" and "I" are shown for adults and children. "A" includes the German phones[1] /a/, /a:/ and the English /a/, /A:/, /V/ whereas "I" includes the phones /I/, /i:/, /j/. The shift into higher frequency regions is well noticeable for both children databases, especially for the first formant. The second formants of the *Youth* database are higher than for *Aibo*.

The variability for the whole frequency range is shown in Fig. 2a , where the averaged (phone independent) speaker vari-

[1] All phone transcriptions are given in the computer-readable phonetic alphabet SAMPA (http://www.phon.ucl.ac.uk/home/sampa/)

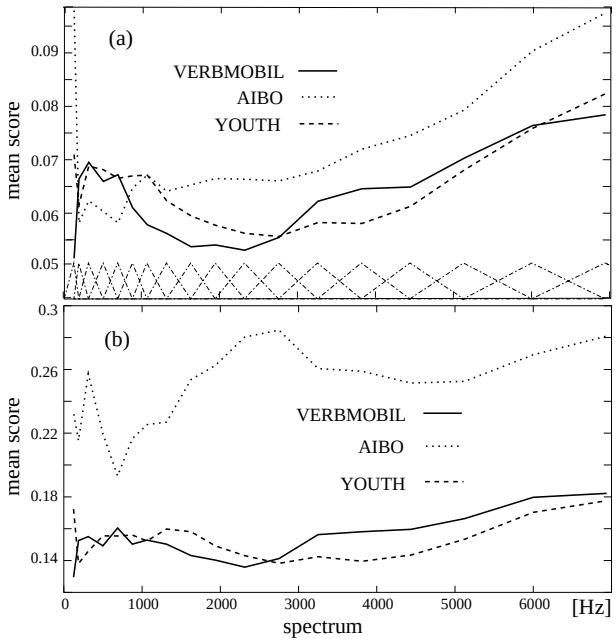

Figure 3: *Scores for the Mel filterbank coefficients (a) and corresponding TRAPs (b) achieved by a Gaussian classifier with 1-dimensional (a) resp. 33-dimensional (b) input.*

ances for 18 Mel spectrum energies are compared. The fricatives cause the increased variances in the high frequency area whereas the formants are responsible for the peak around 2 kHz (*Youth*, *Aibo*) and around 1.6 kHz (*Verbmobil*). Below 3.5 kHz the variabilities for *Youth* are higher than for adults. The *Aibo* database shows smaller variabilities because of its much simpler vocabulary. Note that the peaks at 2.1 kHz appear in both databases and even *Aibo* shows higher values than the adults, whereas the corresponding peak for *Verbmobil* is at a lower frequency. If we analyze the variances separately for male and female speakers, a shift of the female's trajectory to higher frequencies and higher variances may be observed for *Verbmobil* (Fig. 2b, plotted in higher resolution with 30 Mel filters) whereas the children's data is highly correlated (not shown).

Comparable to the trajectories in Fig. 2a is Fig. 3a where the recognition performance for each of the 18 Mel spectrum energies is shown. We train Gaussian classifiers on 1-dimensional features and calculate the class-wise averaged scores for phone supersets (Sec. 3.4). The correlation between the scores and the variances per filterbank is between 0.9 (*Youth*) and 0.6 (*Verbmobil*). In Fig. 3b we apply TRAPs instead of filter energies to the classifier, now the input has dimension 33. As can be seen in the figure, the recognition rate rises in comparison to Fig. 3a, especially for *Aibo*. An additional peak can be found between 1.5 and 3 kHz for *Aibo* and around 1.5 kHz for *Youth*: The recognition in the frequency range of formants increases more clearly for children, if temporal context is taken into account. The correlation between the trajectories in Fig. 3a and Fig. 3b is high but this is not the case for *Aibo* because of the smaller vocabulary. All trajectories are similar for the class wise averaged recognition rate, but the increase is more flat in the high frequency region since only the fricatives reach high scores.

For the lower frequency region can be observed, that for the children's speech bands with high variability extend into higher frequency than for the adults. In these bands the best improvement is achieved, if context is applied. Next, the correlation of mean TRAPs for phone supersets ("I" in Fig. 4) in different corpora is measured. Since the mean of the Mel spectrum coefficients has been subtracted (Sec. 3.1) and since the variance of TRAPs belonging to the same phone is minimal around the center frame, the mean TRAPs have a higher amplitude there. Best correlation between mean TRAPs was observed between *Verbmobil* and *Youth*. However, for vowels frequency bands could be found where the correlation of the children's TRAPs is very high in the average: around 690 Hz and 2.3 kHz.

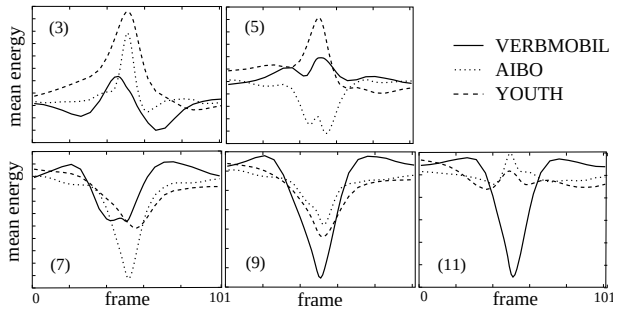

Figure 4: *Mean TRAPs for superset "I" in the Mel filterbanks 3, 5, 7, 9, 11 centered at 312, 687, 1062, 1625, 2312 Hz.*

4.2. Phone classification

In this section we focus on the classification of phone supersets. The class-wise averaged frame recognition rates on the validation data are shown for 24-dimensional standard features (Sec. 3.1) and 12-dimensional TRAP based features (Sec. 3.3) in Tab. 1. For *Aibo* the TRAPs outperform the standard features

	Recognition rate			Complem. info	
	MFCC	TRAP	all	MFCC	TRAP
Verbm.	56.0	50.2	57.8	28.6	20.5
Aibo	57.1	57.8	65.9	21.7	23.1
Youth	54.4	50.4	59.1	31.1	25.6

Table 1: *Phone recognition rate in % for MFCC(+Δ) or TRAPs (left). % of correctly classified frames that are only correct for the respective features (right).*

whereas for the adults' speech database the strongest decrease in recognition rate is observed. Furthermore, there is complementary information in both feature sets: For *Aibo* 21.7 % of frames recognized with cepstral features are not recognized with TRAPs, and vice versa 23.1 % (Tab. 1). For children, the complementary information is particularly high for nasals ($\approx$ 36%). A combination of standard features and TRAPs can improve recognition rates, in particular for the children databases. For *Aibo* the recognition rate rises to 65.9 %.

For the spectrum based features we investigate how the size of the frequency range covered by the filterbank with 22 filters affects recognition. We assume, that the recognition rate of the phone classifier would correlate with the accuracy of a speech recognizer, which will be shown in the next section. The filterbank is squeezed or expanded by choosing different *maxfrequency* values which limit the range covered by the filterbank. Consequence of higher *maxfrequency* values is also a coarser resolution. The optima depend on the phone: for fricatives *maxfrequency* should be 8 kHz, for vowels somewhere below.

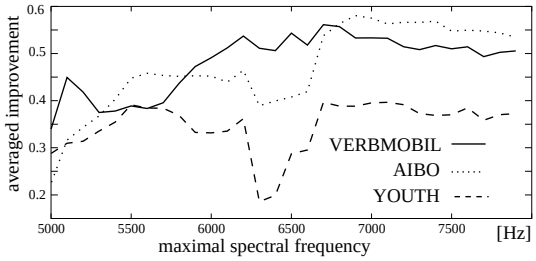

Figure 5: *Averaged improvement achieved by classifiers trained on MFCCs (+Δ) calculated in different spectral ranges covered by the filterbank.*

	Baseline	MFCC	Δ	TRAP
Verbmobil	56.5	36.0	44.9	30.9
Aibo	69.7	52.6	62.1	52.1
Youth	80.7	63.3	70.7	55.7

Table 2: *Word accuracy in % for the baseline system and for feature streams when two of three codebooks are switched off.*

Since the range of recognition performance is rather diverse for different vowels, we measure (class wise averaged) the relative improvement per phone with respect to the interval spanned by the phone dependent minimal and maximal score that could be achieved. The trajectories that are highly correlated for both children corpora are shown in Fig. 5.

4.3. HMM-based word recognition

For the following experiments speech recognizers with a unigram language model are evaluated on the test data set. MFCC are calculated on a filterbank with 22 filters. First, the result from Fig. 5 is corroborated. For *Verbmobil* with *maxfrequency* values of 6250, 7000, and 8000 Hz, a WA of 56.1 %, 56.5 %, 55.8 % is achieved, it drops if frequencies above 7 kHz are covered by the filterbank. For *Youth* we obtain increasing results of 79.3 %, 79.9 %, and 80.7 % WA. The results for *Aibo* with *maxfrequency* values of 6250, 6800, 8000 Hz are 69.4 %, 69.7 %, 69.3 % WA. Tab. 2 shows the optimal baseline results.

If speech recognizers with three codebooks for MFCC, derivatives and TRAPs are estimated, we cannot yet achieve a significant improvement with respect to the standard two codebook recognizer. However, the TRAPs' codebook includes only 24 supervised trained classes, whereas the other codebooks contain 250 classes each. Codebooks can be switched off when corresponding stream weights are set to zero. The results in Tab. 2 show that despite the small codebook TRAP features perform for AIBO as well as MFCC.

5. Conclusion and future work

In this paper, we dealt with spectral features and TRAPs. High correlation between both children's speech databases is observed for the speaker variability per frequency band. Variability and recognition performance are compared. If we consider TRAPs instead of short-time spectrum energies, the increase of recognition performance rises for young speakers particularly in the range of higher formants. In some of those Mel filterbanks the correlation of children's TRAPS proves to be higher, whereas in most cases *Youth* and *Verbmobil* are more similar. *Aibo* has a smaller vocabulary, variabilities are smaller, and with TRAPs the context can be learned better. In phone recognition a combination of MFCC and TRAPs could increase recognition rate, particularly for children. The complementary information in both feature sets is high. We aimed at finding optimal ranges covered by the Mel filterbank. By integration of the TRAPs feature stream into our recognizer, WA did not improve. For the small vocabulary spontaneous speech database *Aibo*, WA of MFCC and TRAP based features is comparable high. In the future, different phone supersets, a restriction to TRAPs in critical bands and a larger band dependent context may improve speech recognition. Further we will explore a corpus, that contains read speech of the same children as in *Aibo*, but with a larger vocabulary.

6. Acknowledgments

This work was partially funded 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.

7. References

[1] A. Potamianos and S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603– 616, 2003.

[2] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. ICASSP*, 1996, pp. 349–352.

[3] D. Giuliani and M. Gerosa, "Investigating recognition of children's speech," in *Proc. ICASSP*, vol. 2, 2003, pp. 137–140.

[4] G. Stemmer, C. Hacker, S. Steidl, and E. Nöth, "Acoustic normalization of children's speech," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 1313–1316.

[5] G. Stemmer, "Modeling variability in speech recognition," Ph.D. dissertation, Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung, Germany, 2004, to appear.

[6] Q. Li and M. Russell, "An analysis of the causes of increased error rates in children's speech recognition," in *Proc. ICSLP*, 2002, pp. 2337–2340.

[7] H. Hermansky and S. Sharma, "Traps - classifiers of temporal patterns," in *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, Sydney, Australia, 1998.

[8] P. Jain and H. Hermansky, "Beyond a single critical-band in trap based asr," in *Proc. Eurospeech*, Geneva, Switzerland, 2003.

[9] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russell, and M. Wong, "'You stupid tin box' - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus," in *Proc. of the 4th International Conference of Language Resources and Evaluation LREC*, Lisbon, 2004, to appear.

[10] G. Stemmer, V. Zeissler, C. Hacker, E. Nöth, and H. Niemann, "A phone recognizer helps to recognize words better," in *Proc. ICASSP*, vol. 1, 2003, pp. 736–739.

[11] C. Hacker, G. Stemmer, S. Steidl, E. Nöth, and H. Niemann, "Various information sources for HMM with weighted multiple codebooks," in *Proc. of the Speech Processing Workshop*, A. Wendemuth, Ed., Magdeburg, Germany, 2003, pp. 9–16.