From Walter G. Kropatsch et al., *Pattern Recognition 27th DAGM Synopsium Vienna, Austria, August/September 2005 Proceedings* , pp. 133–140,

Robust Parallel Speech Recognition in Multiple Energy Bands

Andreas Maier, Christian Hacker, Stefan Steidl, Elmar Nöth, and Heinrich Niemann

Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung, Germany
noeth@informatik.uni-erlangen.de
http://www5.informatik.uni-erlangen.de/

Abstract. In this paper we will investigate the performance of TRAP-features on clean and noisy data. Multiple feature sets are evaluated on a corpus which was recorded in clean and noisy environment. In addition, the clean version was reverberated artificially. The feature sets are assembled from selected energy bands. In this manner multiple recognizers are trained using different energy bands. The outputs of all recognizers are joined with ROVER in order to achieve a single recognition result. This system is compared to a baseline recognizer that uses Mel frequency cepstrum coefficients (MFCC). In this paper we will point out that the use of artificial reverberation leads to more robustness to noise in general. Furthermore most TRAP-based features excel in phone recognition. While MFCC features prove to be better in a matched training/test situation, TRAP-features clearly outperform them in a mismatched training/test situation: When we train on clean data and evaluate on noisy data the word accuracy (WA) can be raised by 173 % relative (from 12.0 % to 32.8 % WA).

1 Introduction

Noise and reverberation have a strong influence on automatic speech recognition systems. Even slight noise or reverberation can cause an enormous decrease of the recognition rate. Human beings are less affected by such disadvantageous effects. Unfortunately these disadvantageous sound situations are quite important in many application scenarios like driving a car or being in an "intelligent room" in which many appliances can be controlled by voice. The use of closetalk microphones is often not practical since they have to be attached very close to the speaker's head and the user acceptance of such a device is very low. This is the reason why robust automatic recognition systems with far distant microphones are desirable. To achieve this we look at two different aspects of automatic speech recognition: The use of features which are restricted to certain frequency bands but are calculated over a longer time span and the use of multiple recognizers.

Our features are based on the TRAP approach of Hynek Hermansky presented in [1]. In his investigations he especially pointed out the robustness of his features.

The speech corpus used in this paper is available in three different levels of noise and reverberation. In addition to the recordings with a closetalk microphone the scenery was filmed by a video camera for documentary purposes. In this manner a second rather noisy version could be recorded. The third version was obtained by adding an artificial reverberation to the closetalk version.

Since the TRAP-features are computed for each band individually it is easy to train different independent recognizers. In case of a distortion of a single frequency band the other recognizers can still recognize the spoken utterance. In [2] it is shown that such a combination using the ROVER procedure can produce very promising results. A recognizer with a certain feature set is trained on every version of the corpus. Each feature set is created using different energy bands. Furthermore different types of TRAP-features are used as well. For the baseline system MFCCs are used.

After the training the recognizers are evaluated on a disjoint test set. We present results on phone and on word level. Furthermore outputs on word level are joined using ROVER included in the Speech Recognition Scoring Toolkit (SCTK) which can be downloaded from [3].

All recognizers trained on the closetalk and closetalk reverberated version of the corpus are evaluated on the test set of the room microphone version as well.

The next section gives a short literature overview. In chapter 3 a short description of the AIBO database follows. The experimental setup and the results are described in chapter 4. The paper ends with an outlook to future work and a summary.

2 Related Work

The feature extraction is primarily based on the work of Hynek Hermansky. In [1] the so called **T**empo**ra**l **P**atterns (TRAP) are introduced. Based on the assumption that the temporal context in each band is essential for the classification of phones each band is analyzed individually, first. Using a long context of up to one second a neural net is trained for each energy band. The outputs of the nets are scores for 29 phonetic classes. These scores are obtained for each band and used as input for a neural net merger. The merger's outputs are again scores for 29 phonetic classes. As Hermansky states these features can produce better results than perceptive linear prediction (PLP) features especially when the distortions occur only in certain bands.

Furthermore Hermansky points out the importance of the modulation spectrum in [4]. This spectrum can be obtained by a short-time analysis of each TRAP band. From this analysis a spectrum results whose both axes have frequency scales. One results from the filter bank analysis and displays the frequency of the signal. The other one displays the modulation frequency. As Her-

mansky states only modulations over 1 Hz and below 16 Hz are important for the perception of speech.

In [5] the importance of the modulation spectrum is highlighted even more. Greenberg proposes to use a so called modulation spectrogram. This kind of spectrogram does not display the strength of a certain frequency band anymore. It displays only the strength of the modulation around 4 Hz. In order to compute this, the spectrogram is processed for each frequency band and transformed into modulation frequency. It is processed in such a manner that in the end only a single coefficient is obtained for the modulation frequency between 2 and 8 Hz. These coefficients get arranged in a spectrographic layout to form the modulation spectrogram. Greenberg states that the modulation spectrogram is quite robust to noise and distortions.

In order to combine the recognizers the ROVER system is employed. [2] describes the algorithm as follows. The output of the recognizers on word level are aligned first and merged later on with the ROVER voting module. In order to do the alignment two hypotheses are processed iteratively by dynamic programming. To match the task the algorithm was extended by Fiscus et al. since the base hypothesis can hold more than one word at a time. Afterwards the best word for each alignment can be found. This is done with the voting module. It supports several modes which can even include the scores returned by the recognizer into the decision process. A reduction of up to 16 % of the word error rate was obtained.

In [6] it is shown, that reverberation can be created artificially if the characteristic spectral properties are known. Those properties can be acquired by recording a known signal from various positions in a room; for each position in the room attributes can be determined. The characteristics found can be applied to a clean signal by convoluting the signal with a finite-impulse-response (FIR) filter. In this manner reverberation can be added to any signal.

3 Corpus

As already mentioned in the introduction the AIBO database is available in three versions. The experimental setup of this database used a Sony AIBO robot [7]. The original design was intended to record emotional speech of children. The children were to accomplish several tasks with AIBO commanding it by voice. However, the robot was controlled by a wizard in a so called *wizard of Oz* experiment. The wizard was disguised as an audio technician. A complete description of the tasks can be found in [8]. The recordings were done with a closetalk microphone which was attached to the child's head. Thus a clean version of the data was recorded which is called closetalk (ct) later on.

For documentary purposes the whole experiment was filmed with a video camera as well. The sound track of the film contains a lot of reverberation and background noises, since the camera's microphone is designed to record the whole scenery in a room and since the camera was approximately 3 m away from the child. The fact that the distance between speaker and microphone was quite far

and that the child was not facing the microphone emphasize the difficulty of this recognition task. This version is called room microphone (rm) in the following.

The third version of the corpus was created using artificial reverberation. To achieve this the data of the ct version were convoluted with different impulse responses. The impulse responses were recored in a different room using multiple speaker positions and echo durations T_{60} as shown in Fig. 1. With each of the twelve responses 1/12 th of the corpus was reverberated. This corpus is called closetalk reverberated (ct rv) later on.

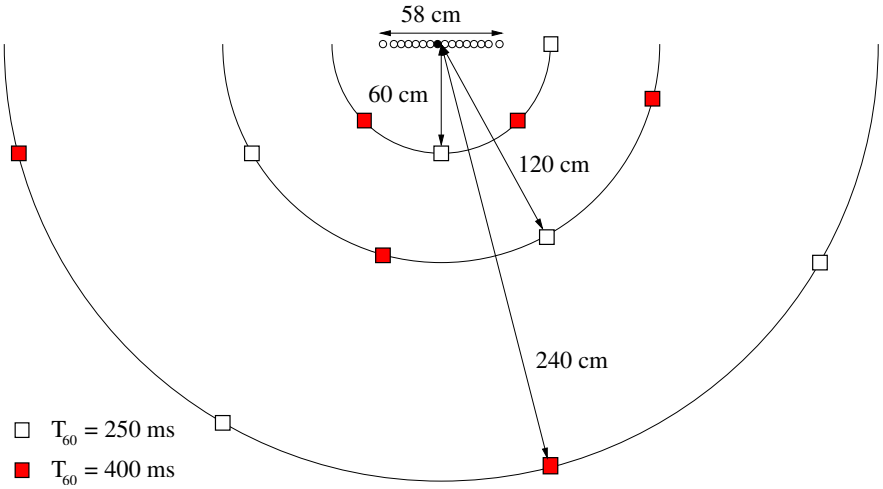

Fig. 1. Positions of the different impulse responses.

The advantage of this procedure is that the data had to be transcribed only once. So all versions have the same size in vocabulary (850 words and 350 word fragments) and the same language models.

4 Experimental Setup

First, we give an overview of the used features, i.e. MFCCs, TRAPs, and filtered TRAPs. Then a description of the recognizers and their combination follows.

4.1 Features

The baseline system which is compared to all results uses MFCC features. So the signal is processed with a fast Hartley transform with a window size of 16 ms computed every 10 ms. Then 22 filter banks are computed from the resulting spectrum. After taking the logarithm the signal is processed with a discrete Cosine transformation. In the end 11 Mel coefficients plus the signal's energy

are taken as static features and another 12 delta features are computed using a regression line over 5 frames, i.e. 56 ms. Cepstrum mean abstraction is applied.

The TRAP features used in this paper are based on [1] but differ in several details. They are computed from the logarithmic Mel spectrum. The Mel filter bank consists of 18 banks. Using a context of ±15 frames a TRAP with 31 entries is created for each band. In this phase n of 18 bands are chosen. The coefficients of these bands are concatenated and reduced to 24 dimensions using a linear discriminant analysis (LDA) instead of the neural nets used by Hermansky. The classes needed by the LDA are obtained by a forced alignment with the baseline hidden Markov recognizer. In total 47 German phonetic classes were used. In the following the TRAP-features are labeled with the letter T.

The modulation spectrum is computed in a similar manner. After the TRAP coefficients are found they are transformed with a fast Fourier transformation. This results in a complex spectrum which states the modulation of the different energy bands. So the filtered TRAP features can be computed using a band pass between 1 Hz and 16 Hz in modulation frequency. All coefficients of the spectrum outside these boundaries are set to 0. Then the modulation spectrum is transformed again into the TRAP domain. Fig. 2 shows the TRAP of the 8th band of a phone /i/ before and after the filtering on the different versions of the corpus. As can be seen the filtered curves are closer to each other, i.e. the filtering reduces the differences in the features caused by the different recording conditions. Again the desired bands are concatenated and reduced with a LDA transformation. This type of feature is called F in the following.

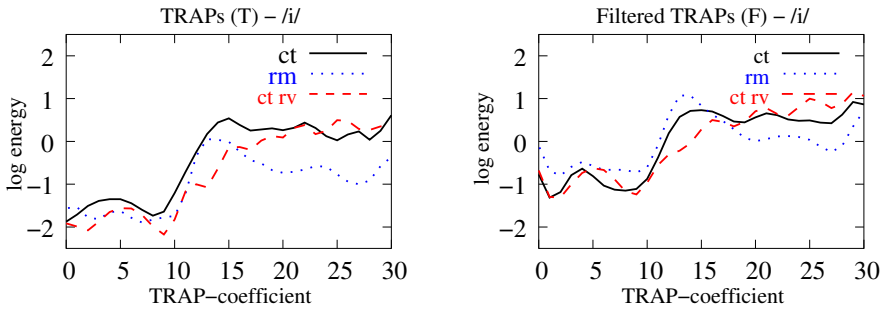

Fig. 2. TRAPs and Filtered TRAPs for the eighth band (from 1.1 kHz to 1.6 kHz) of a phone /i/ for different versions of the corpus.

4.2 Recognizers

The hidden Markov models (HMM) which are employed here are trained with the ISADORA [9] system of the Chair for Pattern Recognition of the University of Erlangen–Nuremberg. This system has been developed since 1978 and has been employed on various tasks in the field of pattern recognition from speech

recognition to genetic decoding and recognition of hand writing. The system provides all tools needed for the training of HMMs. For decoding the *lr_beam* recognizer is employed. In our experiments we used semi-continuous HMMs with full covariance matrices and polyphone models as elementary HMMs. The latest version of the system is described in [9]. Each recognizer provides a best recognized word chain. These word chains are used as input for ROVER. Using the output of multiple recognizers it can compute a single best word chain. To do this the word chains are aligned iteratively first. Then the best word for each position is chosen. This can be done using several algorithms. In our experiments we used voting and included confidence scores from the recognizers (method *avgconf*).

4.3 Results

In the following the presented features are evaluated. All of the recognizers trained here use only 6 of 18 critical bands in order to enable a majority decision with three recognizers. Tab. 1 gives an overview of the results obtained on phone level. As one can see the recognizers using every third band are far better than the baseline MFCC recognizer. However, if the bands are chosen consecutively like in the case of T_{1-6} the recognizers perform worse than the baseline. Therefore the results on word level focus only on the recognizers using every third band.

Table 1. Results of the different features on phone level; class wise averaged recognition rates in %.

Feature	Bands	Abbreviation	ct	ct rv	rm
MFCC			42.3	36.7	28.2
T	1,4,7,10,13,16	T_{3n+1}	**52.2**	**45.2**	**32.2**
T	1,2,3,4,5,6	T_{1-6}	41.2	33.4	22.0
F	1,4,7,10,13,16	F_{3n+1}	49.7	42.9	28.2

On word level the recognition rates of the individual recognizers can not compete with the baseline recognizers in most cases. Only if the training and the test data do not match like in the case of training on closetalk data and evaluation on room microphone data, some of the TRAP recognizers can obtain better results than the baseline system. Note, that the use of artificial reverberation during training always results in an improvement when evaluating on the room microphone test set. In order to focus on the performance of the acoustic models, Tab. 2 shows the recognition rates for a unigram language model.

When the recognizers are joined with ROVER better recognition rates can be achieved. The TRAP based recognizers can now obtain better results than the baseline in most cases. Nevertheless MFCC still return better results when training and test is done on closetalk microphone data. Tab. 3 gives 4-gram recognition rates, since such a language model is usually applied on a real task. Its perplexity on the test set is 50. Note that the consequence of the combination

Table 2. WAs of the individual recognizers with a unigram language model.

Training Test	ct ct	ct rv ct rv	rm rm	ct rm	ct rv rm
MFCC	**69.3**	**63.1**	**35.2**	**4.9**	7.6
T_{3n+1}	55.0	50.5	27.6	-10.8	1.3
T_{3n+2}	53.2	48.8	27.0	-4.0	**9.4**
T_{3n}	55.1	49.1	29.1	-4.4	2.3
F_{3n+1}	57.3	**51.7**	29.1	-2.4	5.5
F_{3n+2}	55.8	50.5	28.5	**5.8**	**12.5**
F_{3n}	**57.6**	51.5	**29.5**	1.2	6.7

of the recognizers is an enormous improvement if training and test data do not match. Except for the ct/ct constellation the TRAP based recognizers give better results than the MFCC recognizers. Combining the TRAP-features with MFCC provides a small further improvement. When training is done on closetalk data and evaluation on room microphone data the recognition rate can be improved by 173 % relatively (from 12.0 % to 32.8 % WA).

Table 3. Categorical 4-gram recognition rates of the combined recognizers.

# of recognizers	Training Test	ct ct	ct rv ct rv	rm rm	ct rm	ct rv rm
1	MFCC	**77.2**	63.1	46.9	12.0	18.8
1	F_{best}	66.1	61.1	39.0	15.8	23.0
3	$T^{\mathbf{3n}}$	71.1	65.9	47.3	21.6	34.3
4	$T^{\mathbf{3n}}$+MFCC	70.3	**66.7**	48.7	**32.8**	31.6
3	$F^{\mathbf{3n}}$	72.1	63.4	48.9	27.2	34.3
4	$F^{\mathbf{3n}}$+MFCC	73.4	64.5	**49.4**	30.9	**35.2**

5 Outlook and Summary

It could be shown that the TRAP features have the greatest advantage when training and test data are mismatched. We conclude that these features generalize better. Furthermore the use of TRAP features and artificial reverberation can improve the recognition rate on the room microphone test set from 12.0 % to 35.2 %. As the experiments show the upper limit (training and test on room microphone data) is 46.9 % in the baseline system. Fig. 3 gives an overview over the best achieved WAs. In our current research we adapt the ct rv recognizer with a small amount of "in task" data, i.e. rm data. We hope that this will close the gap between the currently best "mismatched" recognizer and the rm/rm baseline. Such a procedure will allow to use the vast amount of clean transcribed training data for the training of far distant microphone recognizers.

In this paper we presented the AIBO database in three versions. These versions are closetalk microphone, room microphone, and closetalk reverberated.

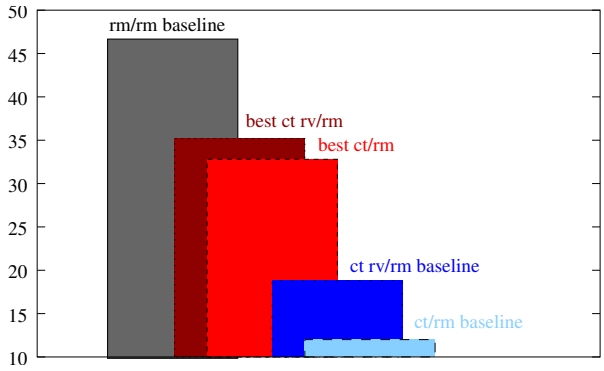

Fig. 3. Overview of the WAs on the room microphone test set.

Furthermore two TRAP based features were introduced. T-features are based on Hermansky's TRAP approach and F-features are created by filtering in the modulation frequency. Then a method to combine multiple recognizers was presented. In the results section the different recognition rates were given. The best improvement compared to an MFCC baseline system could be obtained on mismatched training and test data where a 173 % relative improvement was achieved (from 12.0 % to 32.8 % WA).

References

1. H. Hermansky and S. Sharma. TRAPs - Classifiers of Temporal Patterns. In *Proc. ICSLP '98*, volume 3, pages 1003–1006, Sydney, Australia, 1998.
2. J. Fiscus. A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction. In *Proc. IEEE ASRU Workshop*, pages 347–352, Santa Barbara, USA, 1997.
3. Speech Recognition Scoring Toolkit (SCTK). NIST Spoken Language Technology Evaluation and Utility. http://www.nist.gov/speech/tools/, last visited 28.03.2005.
4. H. Hermansky. The Modulation Spectrum in Automatic Recognition of Speech. In *IEEE Workshop on Automatic Speech Recognition and Understanding*, Santa Barbara, USA, 1997.
5. S. Greenberg and B. E. Kingsbury. The Modulation Spectrogram: In Pursuit of an Invariant Representation of Speech. In *Proc. ICASSP '97*, pages 1647–1650, Munich, Germany, 1997.
6. L. Couvreur and C. Couvreur. On the Use of Artificial Reverberation for ASR in Highly Reverberant Environments. In *Proc. of 2nd IEEE Benelux Signal Processing Symposium*, Hilvaranbeek, The Netherlands, 2000.
7. Sony Europe. AIBO Europe - Official Website, 2004. http://www.aibo-europe.com, last visited 19.12.2004.
8. A. Batliner, C. Hacker, S. Steidl, and E. Nöth. "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 '04*, pages 171–174, Lisbon, Portugal, 2004.
9. G. Stemmer. *Modeling Variability in Speech Recognition*. PhD thesis, Universität Erlangen-Nürnberg, Lehrstuhl für Mustererkennung, Germany, 2005.