GAMMATONE FEATURES AND FEATURE COMBINATION FOR LARGE VOCABULARY SPEECH RECOGNITION

R. Schlüter[1], *I. Bezrukov*[1], *H. Wagner*[2], *H. Ney*[1]

[1]Lehrstuhl für Informatik 6 - Computer Science Department
[2]Lehrstuhl für Biologie II - Biology Department
RWTH Aachen University, Aachen, Germany
schlueter@cs.rwth-aachen.de

ABSTRACT

In this work, an acoustic feature set based on a Gammatone filterbank is introduced for large vocabulary speech recognition. The Gammatone features presented here lead to competitive results on the EPPS English task, and considerable improvements were obtained by subsequent combination to a number of standard acoustic features, i.e. MFCC, PLP, MF-PLP, and VTLN plus voicedness. Best results were obtained when combining Gammatone features to all other features using weighted ROVER, resulting in a relative improvement of about 12% in word error rate compared to the best single feature system. We also found that ROVER gives better results for feature combination than both log-linear model combination and LDA.

Index Terms— feature extraction, auditory systems, gammatone filterbank, acoustic feature combination, speech recognition

1. INTRODUCTION

The starting point of this work was a cooperation between the Computer Science and the Biology Dept. of RWTH Aachen University. The aim was to use biologically inspired acoustic features for speech recognition. In the course of this work a number of biologically inspired features were tested. This included features consisting of auditory filterbanks optionally supplemented by models of the inner hair cells as well as inner hair cell – auditory synapse processing stages. With specific focus being on robustness, a large number of experiments were carried out on the AURORA 2 (and 4) tasks.

In the course of this work, Gammatone (GT) features especially resulted both in improvements for noisy data, and even gave slightly better results on clean data. At this point we started the work presented here. A Gammatone-based feature extraction frontend was integrated into the signal-processing framework of the RWTH large vocabulary speech recognition system, and Gammatone features were tested on a large vocabulary speech recognition task, here the European Parliament Plenary Sessions (EPPS) English task from the TC-STAR project [12]. Since the results were competitive, systematic experiments for combination of Gammatone features with a number of other, state-of-the-art acoustic features were performed, leading relative improvements in word error rate of about 12% compared to the best performing single feature system.

In Sec. 2 the extraction of the Gammatone features used here is described. Sec. 3 introduces the set of acoustic features the Gammatone features were combined to and discusses the combination approaches used here. Finally, in Sec. 4 results for speech recognition experiments with Gammatone features are presented together with a systematic comparison of explicit and implicit feature combination approaches, followed by the conclusions and an outlook on further work in Sec. 5.

2. GAMMATONE FEATURES

In this section we present an acoustic feature extraction based on an auditory filterbank realized by Gammatone filters. The Gammatone filter was introduced in [1]. In [2], Gammatone filters were used for characterizing data obtained by reverse correlation from measurements of auditory nerve responses of cats. The filter is defined in the time domain by the following impulse response:

$$h(t) = k \cdot t^{n-1} \exp(-2\pi \cdot B \cdot t) \cdot \cos(2\pi \cdot f_c \cdot t + \phi).$$

Here, k defines the output gain, B defines the duration of the impulse response and thus the bandwidth, n is the order of the filter and determines the slope at the edges, f_c is the filter's center frequency, and ϕ the phase. For filter orders of $n = 3, \ldots, 5$, the Gammatone filter is reported to give a good approximation of the human auditory filter. In this work, 4th order Gammatone filters were used, implemented as infinite impulse response filters according to [9] and [11].

For a sampling rate of 16kHz, the center frequencies of 68 Gammatone filters were distributed over the frequency range according to the Greenwood function with human parameters [5]. The Greenwood function is defined as follows:

$$\rho_{gw}(x) = A \cdot (10^{a \cdot x} - k) \text{ Hz}$$

For human data, suitable parameters are $A = 165.4$ (to yield frequency in Hz), $a = 2.1$ if x is expressed as a proportion of basilar membrane length and $k = 1$ (for adjusting the lower frequency limit of the human ear).

The absolute values of the Gammatone filter outputs were temporally integrated using a 25 ms wide Hanning window width a 10 ms frame shift [6]. A spectral integration with a 9-channel window and a 4-channel shift followed. Then, (10^{th} root or log) compression was performed, followed by cepstral decorrelation resulting in 16 cepstral coefficients. After cepstral decorrelation, normalization methods were applied, including mean and, optionally, variance normalization.

3. COMBINATION OF MULTIPLE ACOUSTIC FEATURES

In this work, Gammatone features were combined to other acoustic feature sets using explicit and implicit combination methods. Explicit combination was done using Linear Discriminant Analysis (LDA). For LDA, specific attention was directed to shortcomings w.r.t. combination of strongly correlated features, as reported in [10].

Table 1. Corpus statistics for the EPPS English task of the 2006 TC-STAR Evaluation Campaign.

corpus	recording period	speech [h]	# run. words
Train06	May'05-Jan'06	87.5	1,600,000
Dev06	Jun'05	3.2	28,000
Eval06	Sept'05	3.2	30,000

Although these shortcomings could be reduced in this work, the results obtained using LDA for feature combination still are unsatisfactory, as discussed in Sec. 4.4. Implicit feature combination subsequently was performed using log-linear model combination to combine acoustic models trained on individual acoustic feature sets, as well as using ROVER to combine systems built using individual feature sets.

The individual feature sets used for combination experiments with the Gammatone features presented here comprise Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP) features, Mel-Frequency PLP (MF-PLP) features, as well as MFCC-based Vocal Tract Length Normalization based features plus a voicing feature (VTLN-VOI). Details on the implementation of these features used here are given in [13].

4. RESULTS USING MULTIPLE FEATURES

In this section, results for using Gammatone features in ASR and for combination of Gammatone features with state-of-the-art acoustic features are presented.

4.1. EPPS English Corpus

For all the experiments presented here, the European Parliament Plenary Sessions (EPPS) English corpus as defined for the 2006 TC-STAR Evaluation Campaign was used [8] was used. The EPPS corpora were built within the European project *Technology and Corpora for Speech to Speech Translation* (TC-STAR) [4, 12]. The corpus statistics are given in Table 1. The acoustic training was performed on the *Train06* corpus. The *Dev06* corpus was used for parameter optimization, e.g. of the language model scaling factor. The optimized system was then evaluated on the *Eval06* corpus.

4.2. Experimental Setup

All experiments were performed using a common training procedure, for the sake of comparability of the results resulting from the variety of acoustic features. The training was not done from scratch. Instead, an initial alignment created by the MFCC baseline model was used to generate the models in the first iterations. The phonetic decision tree for the first iterations was also taken from the MFCC baseline. It consisted of 4,500 generalized triphone states, plus one for silence. Each state was modeled with a Gaussian mixture distribution with a global pooled covariance matrix.

Altogether three iterations of maximum likelihood training were performed. In the first iteration, the features were augmented with derivatives and no LDA matrix was trained. Single Gaussian densities were estimated using the initial alignment, and 8 splits were performed, resulting in a total number of about 900k densities.

In the second iteration, an alignment was generated using the model from the first iteration. Then, a phonetic decision tree was built, based on the new alignment, followed by the estimation of an LDA matrix. A second phonetic decision tree and a second LDA were estimated afterwards. In the next step, single Gaussian densities were estimated and split 8 times. The LDA transformation was applied to 9 consecutive time frames, and an output dimension of 45 was used. The third iteration was done in the same way as the second. The initial alignment was created using the model from the previous iteration.

A 4-gram language model with modified Kneser-Ney discounting was used for recognition. The language model scaling factor was optimized on the development set.

4.3. Baseline Results: Single Feature Systems

An MFCC frontend with logarithmic compression and mean normalization was taken as baseline. Additionally, systems with 10th root compression and variance normalization were trained to compare the effects of different postprocessing on the MFCC and Gammatone-based features.

The Gammatone feature extraction presented here was compared with the performance of the standard feature extraction methods MFCC, PLP and MF-PLP, the results are shown in Table 2. The best result with Gammatone-based features was 17.9% on the development and 14.5% on the evaluation corpus using 10th root compression and mean & variance normalization. These results are similar to the error rates obtained with the standard methods. The error rate of the Gammatone features on the development set is slightly worse than the result of PLP and MFCC. On the evaluation set, the Gammatone features perform as good as the MFCCs.

It should be noted that variance normalization for both MFCC and GT features with 10^{th} root compression gave improvements, whereas degradations were observed using log compression. It is also interesting to notice that the results for 10^{th} root compression are better than using log compression for both MFCC and GT features, cf. Table 2.

Table 2. Baseline results for single acoustic feature systems on the EPPS 2006 English development and evaluation corpora. Mean normalization was applied in all experiments.

feature	compression	variance norm.	WER [%] dev	WER [%] eval
MFCC	log	no	17.5	14.9
		yes	17.9	15.0
	10^{th} root	no	17.7	15.0
		yes	17.5	14.4
GT	log	no	18.3	14.6
		yes	18.9	15.8
	10^{th} root	no	19.2	15.4
		yes	17.9	14.5
PLP	3^{rd} root	no	17.6	14.7
MF-PLP		no	18.4	15.5

4.4. Feature Combination: Linear Discriminant Analysis (LDA)

As discussed in previous work [10], using LDA for feature combination can lead to considerable degradations when combining strongly correlated or even dependent features. The same we observed when combining Gammatone features with 10th root compression and mean & variance normalization with MFCC features with log compression and mean normalization. The results are given in the first row of Table 3. Since both features contain an energy coefficient, in the next step we tried to use only one of the energy coefficients. The

Table 3. Results for LDA-based feature combination of MFCC (log compression and mean normalization) with Gammatone features (10th root compression with mean & variance normalization.

LDA output dimension	cepstral energy from	WER [%] dev	eval
45	GT & MFCC	18.0	14.7
	MFCC only	17.7	14.3
60	GT &MFCC	17.7	14.0
	MFCC only	17.0	14.0

energy coefficients can be assumed to be dependent, and dependency was shown to be a problem for LDA estimation in [10]. As shown in the second row of Table 3, this step leads to an improvement on the evaluation set, but on the development set results are still worse than the better of the two single features (cf. Table 2 for the single feature results). Assuming that the LDA-estimation is problematic for this case, another idea was that the information extracted by LDA is spread upon more output dimension than in the well-estimated case. Therefore, the LDA output dimension was increased from 45 to 60. As shown in Table 3, this step leads to improvements for both the dev and the eval set, provided the energy coefficient is taken from one feature set only. When repeating the original combination experiment for combination of MFCC (with log compression), MF-PLP and PLP features as reported in [10], the latter observation could not be confirmed when using an LDA output dimension of 60. Table 4 shows, that even in the case of using only one of the energy coefficients of all three features, the LDA combination results is still worse than the result of at least the best individual feature based system.

Table 4. Results for LDA-based feature combination of MFCC (with log-compression), PLP, MF-PLP (all with mean normalization), and Gammatone features (10th root compression, mean & variance normalization). The LDA output dimension was 60.

cepstral energy from	WER [%] dev	eval
all features	18.3	15.5
MFCC only	17.6	15.1

4.5. Model Combination: Log-Linear

Due to the shortcomings of LDA for the case of acoustic feature combination, in the next step we investigated log-linear acoustic model combination for the combination of the MFCC (with log compression and mean normalization) and the Gammatone feature set (with 10^{th} root compression and mean & variance normalization). The optimal weight exponent λ is determined by grid search on the development set. The results are given in Table 5. The best result on the development set was obtained with a weight of $\lambda = 0.6$ for the MFCC model and a weight of $1 - \lambda$ for the Gammatone model resulting in a WER of 16.6% on the dev set. The combination results in an absolute improvement of 0.5% on the evaluation corpus, which nevertheless is not better than the corresponding result using LDA for feature combination, cf. Table 3.

4.6. System Combination: ROVER

Finally, ROVER [3] was investigated for the combination of systems based on individual feature sets. Altogether, five acoustic features/systems were used in the ROVER experiments. Besides the systems trained during this work, the output of a system with vocal tract length normalization and a voicing feature is included. This system was used as baseline for the RWTH system in the TC-Star evaluation 2006 [8].

Table 5. Results for log-linear model combination.

System	WER [%] dev	eval
MFCC-LOG-MN	17.5	14.9
GT-10th-MVN	17.9	14.5
MFCC-LOG-MN + GT-10th-MVN	16.6	14.0

The single features system and ROVER combination results are summarized in Table 6. In addition to the standard ROVER approach, we also applied weighted ROVER [7], where prior weights for the individual systems are trained in addition to using confidences. To get an idea of the system combination potential, also oracle error rates were included which represent the best word error rate to be obtained given the ROVER alignment. Here, we investigated the effect of combining Gammatone features to all the other features. For a given number of systems combined, the results are ordered in descending order with respect to single feature system performance, and results for lower numbers of systems combined are given higher up in Table 6. It should be noted that the results

Table 6. Results on the EPPS 2006 English corpus using standard and weighted ROVER. Features used: VTLN-Voicing (VTLN-VOI), MFCC with log compression and mean normalization (MFCC), PLP with log compression and mean normalization (PLP), MF-PLP with log compression and mean normalization (MF-PLP), Gammatone with 10th root compression, mean & variance normalization (GT-10th). System combination: standard ROVER with confidence scores (Standard), and weighted ROVER (Weighted).

GT-10th	MF-PLP	PLP	MFCC	VTLN-VOI	WER [%] Standard dev	Standard eval	Weighted dev	Weighted eval	Oracle WER [%] dev	Oracle eval
X					17.9	14.5	17.9	14.5		
	X				18.4	15.5	18.4	15.5		
		X			17.6	14.7	17.6	14.7		
			X		17.5	14.9	17.5	14.9		
				X	17.0	14.0	17.0	14.0		
X	X				16.6	13.7	16.5	13.8	13.4	11.2
X		X			16.4	13.6	16.3	13.6	13.3	11.2
X			X		16.4	13.6	16.3	13.7	13.1	11.2
X				X	15.8	12.9	15.7	12.9	12.4	10.0
X	X	X			15.9	13.2	15.9	13.2	11.8	9.9
X	X		X		15.7	13.3	15.7	13.3	11.7	9.9
X		X	X		15.7	13.3	15.7	13.3	11.5	9.8
X	X			X	15.3	12.6	15.2	12.5	11.0	9.0
X		X		X	15.3	12.6	15.1	12.5	10.8	8.9
X			X	X	15.4	12.6	15.2	12.6	10.8	9.1
X	X	X	X		15.6	13.0	15.5	13.0	10.8	9.1
X	X	X		X	15.2	12.5	15.0	12.4	10.2	8.4
X	X		X	X	15.3	12.6	15.1	12.4	10.2	8.4
X		X	X	X	14.9	12.5	14.9	12.4	10.0	8.3
X	X	X	X	X	15.1	12.5	14.8	12.4	9.6	7.9

obtained by ROVER are fully consistent, i.e. the error rates decrease both when combining better systems and when increasing the number of systems combined. When combining all features, a relative improvement of approx. 12% in word error rate compared to the best individual feature system is obtained for weighted ROVER.

4.7. Comparison of Feature Combination Methods

In Table 7 the results for combining MFCC and Gammatone features using explicit feature combination with LDA and implicit feature combination using either log-linear model combination or ROVER are presented. Clearly, the best results are obtained using ROVER, resulting in a relative improvement of about 6% in word error rate compared to the better single feature system.

Table 7. Comparison of explicit feature combination using LDA, and implicit feature combination by log-linear model combination and system combination based on acoustic models/systems trained on the individual acoustic features. Here, MFCC and Gammatone features were combined.

single feature systems	WER [%]	
	dev	eval
MFCC-LOG-MN	17.5	14.9
GT-10th-MVN	17.9	14.5
combination methods	dev	eval
LDA (output dim.: 60)	17.0	14.0
log-linear	16.6	14.0
ROVER using confidences	16.4	13.6

5. CONCLUSIONS AND OUTLOOK

In this work we could show that Gammatone features lead to competitive results for large vocabulary speech recognition. Furthermore, different methods to combine Gammatone features with a number of standard acoustic features, i.e. MFCC, PLP, MF-PLP, and VTLN plus voicedness, were investigated. Best results were obtained when combining all features using weighted ROVER, resulting in a relative improvement of about 12% in word error rate compared to the best single feature system. We also found that ROVER gives better results for feature combination than both log-linear model combination and LDA. Although some shortcomings of LDA in case of feature combination could be reduced, LDA still is suboptimal. The latter still is unsatisfying since training and recognition in case of explicit feature combination would be computationally more efficient than implicit feature combination methods like model or system combination, since these require training of, and recognition/scoring using individual models/systems for each set of features. Therefore, in future work we will concentrate on improving efficiency by finding better methods of implicit/explicit feature combination. In addition we will investigate methods to do VTLN using Gammatone features. Further, these studies will be continued using full systems including VTLN, speaker adaptation, speaker adaptive training, and discriminative training. Finally, it should be noted that the oracle error rates presented in Table 6 still are much better than the best combination results, i.e. the potential of system combination by far is not fully exploited, yet. Therefore, further research into improved model and/or system combination methods is due.

Acknowledgements
This work was partly funded by the European Commission under the project TC-STAR (FP6-506738).

6. REFERENCES

[1] A. M. H. J. Aertsen, P. I. M. Johannesma, D. J. Hermes: "Spectro-Temporal Receptive Fields of Auditory Neurons in the Grassfrog," *Biological Cybernetics*, Vol. 38, No. 4, pp. 235–248, Nov. 1980.

[2] E. de Boer, H. R. de Jongh: "On Cochlear Encoding: Potentialities and Limitations of the Reverse-Correlation Technique," *The Journal of the Acoustical Society of America*, Vol. 63, No. 1, pp. 115–135, Jan. 1978.

[3] J. G. Fiscus: "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction," in *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop* (ASRU), pp. 347–352, Santa Barbara, CA, USA, Dec. 1997.

[4] C. Gollan, M. Bisani, S. Kanthak, R. Schlüter, and H. Ney: "Cross Domain Automatic Transcription on the TC-STAR EPPS Corpus," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP), Vol. I, pp. 825–828, Philadelphia, PA, March, 2005.

[5] D. D. Greenwood: "A Cochlear Frequency-Position Function for Several Species - 29 Years Later," *The Journal of the Acoustical Society of America*, Vol. 87, No. 6, pp. 2592–2605, 1990.

[6] W. Hemmert, M. Holmberg, D. Gelbart: "Auditory-based automatic speech recognition," in *Proc. ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, Jeju Island, Korea, Oct. 2004.

[7] B. Hoffmeister, T. Klein, R. Schlüter, H. Ney: "Frame-Based System Combination and a Comparison with Weighted ROVER and CNC," in *Proceedings of the International Conference on Spoken Language Processing* (ICSLP/Interspeech), pp. 537–540, Pittsburgh, PA, September 2006.

[8] J. Lööf, M. Bisani, C. Gollan, G. Heigold, B. Hoffmeister, C. Plahl, R. Schlüter, H. Ney: "The 2006 RWTH Parliamentary Speeches Transcription System," in *Proceedings of the International Conference on Spoken Language Processing* (ICSLP/Interspeech), pp. 105–108, Pittsburgh, PA, September 2006.

[9] E. Lopez-Poveda, R. Meddis: "A Human Nonlinear Cochlear Filterbank," *The Journal of the Acoustical Society of America*, Vol. 110, No. 6, pp. 3107–3118, December 2001.

[10] R. Schlüter, A. Zolnay, H. Ney: "Feature Combination using Linear Discriminant Analysis and its Pitfalls," in *Proceedings of the International Conference on Spoken Language Processing* (ICSLP/Interspeech), pp. 345–348, Pittsburgh, PA, September 2006.

[11] M. Slaney: "An Efficient Implementation of the Patterson-Holdsworth Auditory Filterbank," *Technical Report 35*, Apple Computer Co., 1993.

[12] *Technology and Corpora for Speech to Speech Translation (TC-STAR)*, Integrated Project funded by the European Commission, Project No. FP6-506738, 2004-2007. `http://www.tc-star.org`.

[13] A. Zolnay, R. Schlüter, H. Ney: "Acoustic Feature Combination for Robust Speech Recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP), Vol. 1, pp. 457-460, Philadelphia, PA, March 2005.