AN INVESTIGATION OF CEPSTRAL PARAMETERISATIONS FOR LARGE VOCABULARY SPEECH RECOGNITION

Reinhold Haeb–Umbach, Marco Loog

Philips Research Laboratories
Weisshausstrasse 2, D-52066 Aachen, Germany
{haeb,loog}@pfa.research.philips.com

6th European Conference on Speech Communication and Technology (EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ABSTRACT

We examined variants of MFCC and PLP cepstral parameterisations in the context of large vocabulary continuous speech recognition under different acoustical environmental conditions: Compared to MFCC, mel-frequency PLP uses a cubic root intensity-to-loudness law, and an LPC analysis is applied to the mel-warped spectrum. In LPC-smoothed MFCC, the only difference to MFCC is the additional LPC smoothing of the warped spectrum. While neither technique was able to significantly outperform the MFCC parameterisation in our setup which includes an LDA feature transformation, feature set combination via DMC at the acoustic likelihood level and via ROVER at the recognized word level delivered small but consistent improvements.

1. INTRODUCTION

Parameterisation of an analog speech signal is the first stage in the speech recognition process. Finding a robust speech representation is a precondition for the success of the subsequent recognition steps. Mel-frequency cepstral coefficients (MFCC) [1] are probably the most popular speech feature set which have even been subject to standardisation in the AURORA project [2]. Nevertheless, there is still active research in superior speech representations for speech recognition. A lot of effort is devoted to exploiting physiological and psychoacoustic findings about human perception. Examples are Perceptual Linear Predictive (PLP) analysis [3] and the Ensemble-Interval Histogram (EIH) [4]. See e.g. [5] for a short overview of auditory feature extraction research. Auditory models tend to be computationally more complex than standard feature extraction techniques. This may be one reason why they have not been used extensively in large vocabulary continuous speech recognition. However, some of the ideas, notably those of PLP, have lead to variants of MFCC feature extraction which are used by a couple of groups participating in the Hub-4 evaluations [6].

In this paper we investigate different variants of MFCC and PLP cepstral parameterisations in the context of large vocabulary continuous speech recognition (on Wallstreet Journal (WSJ) and Hub-4 databases). The goal of this study is to gain insight in the importance of different processing steps in the cepstral parameterisations and, ultimately, to improve the widely used MFCC representation. We defined three test scenarios, which we assumed typical for many real-world applications: "matched-clean" (training and test on WSJ data), "matched-noisy" (training and test on Hub-4 data), and "mismatch" (training on WSJ, test on Hub-4).

Further, we investigated options for combining feature sets: a combination at the feature vector level via linear discriminant analysis (LDA), a combination at the acoustic likelihood level via Discriminative Model Combination (DMC) [7], and a combination at the recognized word level via ROVER [8].

The paper is organized as follows: In section 2 the different cepstral parameterisations investigated are described. Section 3 presents a performance comparison of feature sets on the forementioned test scenarios, and section 4 highlights our efforts in improving error rate by combining feature sets. Finally, section 5 contains some conclusions.

2. CEPSTRAL PARAMETERISATIONS

Hermansky has extended Linear Prediction analysis to "Perceptual" Linear Prediction by introducing concepts from psychophysics:

- a "human-like" nonlinear frequency resolution by using the bark frequency scale and trapezoidal critical band filters.
- an approximation of the nonequal sensitivity of the human hearing at different frequencies through an "equal-loudness" preemphasis.
- a nonlinear sound intensity compression by an "intensity-to-loudness" transformation.

Recently people have introduced similar psychophysical concepts into the well-known Mel-frequency cepstral analysis of speech [10], and devised a variant of MFCC called "Mel-frequency PLP" (MF-PLP), see Fig. 1:

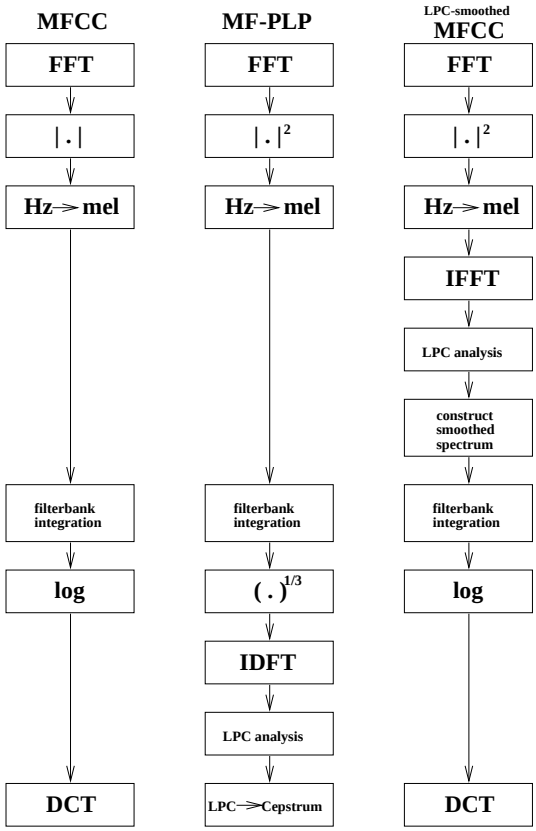

Figure 1: Cepstral parameterisations.

- Rather than using critical bands and a bark scale, the mel-frequency scale and the bank of triangular bandpass filters known from MFCC analysis is used. Actually the difference between the Mel and the Bark frequency scale is very small [11].
- Equal-loudness preemphasis is not included, since the standard preemphasis applied to the speech waveform has a similar effect.
- The log-function present in MFCC is replaced by the cubic root intensity-to-loudness law used in PLP.
- LPC analysis is conducted as in (P)LP analysis. This results in a smoothed spectrum.
- Cepstrum coefficients are computed as is also today common practice in PLP analysis.

Using MF-PLP people reported improved word error rate performance compared to MFCC, notably in training/test mismatch situations [10]. In order to understand the improved robustness we experimented with different configurations, where we subsequently exchanged building blocks of MFCC analysis with corresponding blocks of MF-PLP analysis. Of those, the most promising setup was what we called "LPC-smoothed MFCC", where only the LPC smoothing of the spectrum is added to the MFCC analysis, see Fig. 1. This configuration is similar to what BBN used in their Hub-4 system [12].

3. TEST ENVIRONMENT AND EXPERIMENTAL RESULTS

We defined three test scenarios, representative of a wide range of recognition setups:

- **"matched-clean:"** Training on WSJ0 42 male speakers (7.5h). Test on the four WSJ 5k test sets dev/eval 92/93 (20 male speakers, 13113 words), bigram language model.
- **"matched-noisy:"** gender-dependent training on 96h of Hub-4 training data. Test on Hub-4 eval'97 test set, partitioned evaluation (approx. 3h, 32832 words), trigram language model.
- **"mismatch:"** gender-dependent training on WSJ0+1 database (142 male and 142 female speakers, approx. 40h per gender). Test on Hub-4 eval'96 test set, partitioned evaluation (approx. 2h, 20318 words), bigram language model.

While the WSJ data are read speech in a clean acoustic environment, the Hub-4 data comprise various acoustic conditions and speaking modes, such as clean, prepared speech (F0), spontaneous speech (F1), speech recorded over telephone channels (F2), speech in the presence of background music (F3), speech under degraded acoustic conditions (F4), non-native speakers (F5) and others (FX). The acronyms in parenthesis are the socalled focus conditions, which NIST has used to categorize the data.
In our recognition experiments the incoming speech signal is blocked every 10ms into frames of 25ms width, irrespective of the type of cepstral parameterisation used. Sentence based (in the case of WSJ) or segment based (in the case of Hub-4) cepstral mean normalization is applied to the cepstral feature vectors. Further, in the case of the matched-noisy scenario, the variance per segment is normalized to unity. Rather than computing first and second order time derivatives explicitly, 7 adjacent (static) feature vectors are concatenated to form a large vector which is transformed by a Linear Discriminant Analysis transformation to a 35-component output feature vector.
In the acoustic modeling we employ within-word triphone models and decision-tree clustering. Note that mixtures of Laplacian densities with a single globally pooled mean absolute deviation vector are used.
Table 1 presents the results for the three cepstral parameterisations introduced in section 2 and for the test scenarios described above. Tables 2 and 3 allow a closer look at the results on the Hub-4 databases by showing error rates for the different focus conditions. In the case of MF-PLP we used an LPC analysis of

order 15 while an order of 20 turned out to be optimal for LPC-smoothed MFCC. As can be seen neither MF-PLP nor LPC-smoothed MFCC was able to consistently and significantly outperform the MFCC feature set. In particular there was no overall performance gain by MF-PLP in the mismatch scenario[1].

A Matched-pairs test [9] was carried out between any two of the recognition outputs on the Hub-4 eval'97 test set ("matched-noisy"). Comparing MFCC with MF-PLP led to a P-value of 0.033, which means that under the hypothesis H_0 that the two systems perform equally well, the observed difference would arise in about 3.3% of the occasions. Hence, it is reasonable to regard MFCC and MF-PLP as different. Both other comparisons resulted in P-values larger than 0.14 and so H_0 cannot be rejected.

Table 1: Word error rate in % for MFCC, MF-PLP and LPC-smoothed MFCC feature vectors on 3 test scenarios.

Parameterisation	matched clean	matched noisy	mismatch
MFCC	10.1	21.6	41.8
MF-PLP	10.3	22.1	42.2
LPC-smoothed MFCC	10.1	21.9	41.3
95% conf. interval for MFCC	(9.6, 10.6)	(21.2, 22.0)	(41.1, 42.5)

Table 2: Word error rate in % for MFCC, MF-PLP and LPC-smoothed MFCC feature vectors on matched-noisy test scenarios. Recognition results on Hub-4 eval'97 test data per focus condition.

	All	F0	F1	F2
MFCC	21.6	13.1	20.1	32.2
MF-PLP	22.1	13.4	21.3	31.4
LPC-smoothed MFCC	21.9	13.4	20.5	31.9
	F3	F4	F5	FX
MFCC	30.9	25.6	23.9	37.2
MF-PLP	32.5	26.0	27.1	38.1
LPC-smoothed MFCC	31.4	25.2	25.6	38.8

It is interesting to note that Linear Discriminant Analysis, which is a specialty of our recognition system, reduces the "difference" between the three feature sets and thus the chance to improve beyond

[1] In [13] it was reported that the superiority of MF-PLP only occured after MLLR adaptation while MFCC was better before adaption. In our environment MFCC outperformed MF-PLP even after adaptation (20.0% vs 20.4% in the matched-noisy scenario).

Table 3: Word error rate in % for MFCC, MF-PLP and LPC-smoothed MFCC feature vectors on mismatch test scenarios. Recognition results on Hub-4 eval'96 test data per focus condition.

	All	F0	F1	F2
MFCC	41.8	30.8	37.4	56.8
MF-PLP	42.2	31.6	38.0	58.0
LPC-smoothed MFCC	41.3	30.6	37.4	55.3
	F3	F4	F5	FX
MFCC	46.2	51.8	44.1	60.1
MF-PLP	45.7	50.1	40.1	61.2
LPC-smoothed MFCC	45.5	50.4	45.1	59.5

MFCC. Fig. 2 shows the value of the correlation coefficient ρ_i between the i-th MFCC and MF-PLP feature vector component before and after LDA transformation. Note that in our setup the MFCC signal analysis results in 16 static cepstral coefficients while 20 coefficients have been computed in MF-PLP. In the figure, the 35 features after LDA have been ordered according to decreasing correlation coefficient.

If $|\rho_i|$ were unity, then the MFCC feature x_i could be written as $x_i = ay_i + b$ where y_i is the MF-PLP feature and a, b are constants. Such a linear transformation between input features is however absorbed by LDA, since the optimization criterion of LDA, the trace of the inverse within-class scatter matrix times the between-class scatter matrix, is invariant to linear transformations. Clearly: Let $z = A^T x$ be the transformed MFCC feature vector and $\tilde{z} = \tilde{A}^T y$ the transformed MF-PLP vector, where A and $\tilde{A}$ are the corresponding LDA transformation matrices obtained by maximizing the trace criterion on the corresponding feature set. Then $z = \tilde{z}$ [14]. Full correlation of MF-PLP and MFCC feature vectors would result in identical feature vectors after LDA! Indeed one can see that the feature vectors are less "different" after LDA: the correlation coefficients are larger.

4. FEATURE SET COMBINATION

Neither MF-PLP nor LPC-smoothed-MFCC was able to outperform MFCC consistently in the experiments reported in the last section. Although the correlation analysis revealed a high correlation between the features of the different parameterisations, an "oracle-experiment" on the Hub-4 eval'97 results delivered an error rate of 16.6%; i.e. if we had a wizard which selected the correct word, if present among the three recognition alternatives, we were able to improve from 21.6% to 16.6%! Thus there is quite some room to improve performance by feature set combination. In particular we tried combinations at the feature level, at the acoustic likelihood level and at the recognized word level.

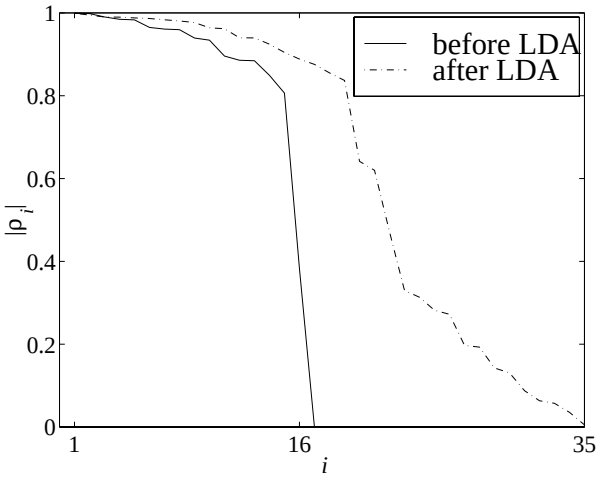

Figure 2: Ordered correlations. Before and after LDA.

A combination at the feature level was conducted by adjoining a MFCC and a MF-PLP vector to one feature vector prior to LDA. The LDA transformation should then deliver the best linear combination of the two cepstral parameterisations. This approach, however, delivered no error rate improvement and was thus abandoned.

A combination at the acoustic likelihood level was achieved via Discriminative Model Combination [7]. DMC aims at an optimal log-linear combination of given acoustic (and, possibly, language) models into one posterior probability distribution. The model weights are trained using discriminative training criteria. We applied DMC to combine word scores of acoustic models trained on the three feature sets in a word lattice.

ROVER [8] is used to combine the recognized word sequences obtained from the three parameterisations. Since we did not use confidence information ROVER amounts to a simple majority voting among the three recognition alternatives.

Table 4 shows that both DMC and ROVER achieve a small error rate improvement on the Hub-4 eval'97 and eval'96 data, compared to the best single feature set, MFCC.

Table 4: Combination of MFCC, MF-PLP and LPC-smoothed MFCC feature sets. Matched noisy scenario. Tests on Hub4-4 eval'97 and eval'96, file 4.

Combination via	matched noisy	
	Hub-4 eval'97	Hub-4 eval'96, file 4
ROVER	21.4	25.6
DMC	21.4	25.5
"oracle"	16.6	20.3
MFCC features	21.6	26.2

5. CONCLUSIONS

The feature extraction variants we experimented with were not able to outperform MFCC consistently in a large vocabulary continuous speech recognition setup, even under training/test mismatch conditions. More research is needed to understand why and how psychophysically motivated processing steps can improve robustness. Minor, though consistent improvements were obtained by feature set combination, both by DMC and ROVER. Compared to what an ideal combination would be able to deliver, the improvement, however, was moderate.

6. REFERENCES

[1] S.B. Davis, P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE T-ASSP*, Vol. ASSP-28, No.4, pp 357–366, Aug. 1980.

[2] URL: http://www-nrc.nokia.com/robust99/DSR/dsr.

[3] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech", *Journal of the Acoustical Society of America*, Vol. 87, pp 1738-1752.

[4] O. Ghitza, "Speech Analysis/Synthesis Based on Matching the Synthesized and the Original Representations in the Auditory Nerve Level", *Proc. ICASSP*, pp 1995-1998, Tokyo, 1986.

[5] N. Kumar, "Investigation of Silicon Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition", Ph.D Thesis, Johns Hopkins University, 1997.

[6] Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop (BNTUW), Lansdowne, VA, Feb. 1998.

[7] P. Beyerlein, "Discriminative Model Combination", *in Proc. ICASSP*, pp 481-484, Seattle, WA, May 1998.

[8] J. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)", *Proc. 1997 IEEE ASRU Workshop*, Santa Barbara, pp 347-354, Dec. 1997.

[9] L. Gillick, S.J. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms", *Proc. ICASSP*, Glasgow, pp 532-535, May 1989.

[10] P.C. Woodland, M.J.F. Gales, D. Pye, S.J. Young, "Broadcast News Transcription Using HTK", *Proc. ICASSP*, Munich, pp 719-722, April 1997.

[11] E. Zwicker, Psychoakustik, (in German), 1982.

[12] F. Kubala et al., "The 1997 BBN BYBLOS System Applied to Broadcast News Transcription", *Proc. DARPA BNTUW*, Lansdowne, VA, Feb. 1998.

[13] S. Wegmann et al., "Dragon System's 1997 Broadcast News Transcription System", *Proc. DARPA BNTUW*, Lansdowne, VA, Feb. 1998.

[14] K. Fukunaga, "Statistical Pattern Recognition", Academic Press, 1990.