Revising Perceptual Linear Prediction (PLP)

Florian Hönig[1], *Georg Stemmer*[2], *Christian Hacker*[1], *Fabio Brugnara*[2]

[1] Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Germany
[2] ITC-irst – Centro per la Ricerca Scientifica e Tecnologica, Povo di Trento, Italy
`hoenig@informatik.uni-erlangen.de, stemmer@itc.it`

Abstract

Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP) are the most popular acoustic features used in speech recognition. Often it depends on the task, which of the two methods leads to a better performance. In this work we develop acoustic features that combine the advantages of MFCC and PLP. Based on the observation that the techniques have many similarities, we revise the processing steps of PLP. In particular, the filter-bank, the equal-loudness pre-emphasis and the input for the linear prediction are improved. It is shown for a broadcast news transcription task and a corpus of children's speech that the new variant of PLP performs better than both MFCC and conventional PLP for a wide range of clean and noisy acoustic conditions.

1. Introduction

The acoustic features most commonly used in speech recognition are Mel Frequency Cepstral Coefficients (MFCC) [1] and Perceptual Linear Prediction (PLP) features [2]. PLP features are reported (e. g. in [3]) to be more robust when there is an acoustic mismatch between training and test data. In our own experiments we found that under clean conditions and when there is no significant mismatch, MFCC features lead to a performance that is slightly superior to PLP. Thus, it is advisable to decide for each task which one of the two feature types would be more appropriate. However, in many applications the acoustic conditions do not remain constant over the whole data set: in broadcast news transcription, for instance, segments with clean speech are intermixed with segments that contain background music or noisy telephone speech. In order to achieve optimal performance it is desirable to have a feature extraction that is well-suited both for clean and adverse acoustic conditions. Thus, the favorable properties of PLP and MFCC have to be combined.

In spite of the fact that PLP has been derived independently of the MFCC technique, there are many similarities between the two methods [4, p. 67]. We utilize these in a revised feature extraction algorithm that integrates elements taken from the MFCC procedure into PLP. Even though PLP has been developed based on a psycho-physical findings, we follow the approach of Hunt [5] and interpret the steps of PLP purely in signal processing terms. We consider the following changes of PLP: (i) the Bark filter-bank is replaced by a Mel filter-bank; (ii) the equal-loudness weighting of the spectrum is substituted by a pre-emphasis that is applied to the speech signal; (iii) the duplication of the first and last filter-bank value that is done in conventional PLP before linear prediction (LP) is dropped. Finally, (iv), a new filter-bank is introduced with a very large number of filters. As the new filters have the same band-width as in the conventional approach, there is no loss of robustness. On a large corpus of broadcast-news data and a corpus of children's speech we show that our variant of PLP, which retains from the auditorily motivated components only the filter-bank, performs better than the conventional methods.

Variants of PLP, in particular the use of alternative filter-banks, have already been investigated by others. Mason and Gu [6] incorporated the JSRU filter-bank in PLP with a slight decrease in performance. Furthermore, the authors claimed that the equal-loudness pre-emphasis had no measurable effect. Woodland et al. [7] proposed MF-PLP, which substitutes the Bark filter-bank in PLP by a Mel filter-bank. This is equivalent to one of the modifications that are also described here. While the authors showed in [7] that MF-PLP compares favorably to MFCC under difficult acoustic conditions, they did not publish any direct comparison between PLP and MF-PLP which is provided in this paper. Furthermore, our approach includes additional improvements of PLP and we show that even the conventional Mel filter-bank can be improved.

This paper is structured as follows: After discussing the similarities and differences between PLP and MFCC, we derive several modifications of PLP. Next, the employed speech corpora and baseline systems are described. Finally, the revised PLP procedure is evaluated.

2. PLP vs. MFCC

In order to justify the proposed modifications, we shortly review PLP and compare it with the MFCC computation. As shown in Fig. 1, PLP consists of the following steps: (i) The power spectrum is computed from the windowed speech signal. (ii) A frequency warping into the Bark scale is applied. (iii) The auditorily warped spectrum is convoluted with the power spectrum of the simulated critical-band masking curve to simulate the critical-band integration of human hearing. (iv) The smoothed spectrum is down-sampled at intervals of ≈ 1 Bark. The three steps frequency warping, smoothing and sampling (ii-iv) are integrated into a single filter-bank called *Bark filter-bank*. (v) An equal-loudness pre-emphasis weights the filter-bank outputs to simulate the sensitivity of hearing. (vi) The equalized values are transformed according to the power law of Stevens by raising each to the power of 0.33. The resulting auditorily warped line spectrum is further processed by (vii) linear prediction (LP). Precisely speaking, applying LP to the auditorily warped line spectrum means that we compute the predictor coefficients of a (hypothetical) signal that has this warped spectrum as a power spectrum. Finally, (viii), cepstral coefficients are obtained from the predictor coefficients by a recursion that is equivalent to the logarithm of the model spectrum followed by an inverse Fourier transform. Fig. 1 shows a comparative scheme of PLP and MFCC computation. Note that the MFCC computation in the baseline-system includes a pre-emphasis $x'_t = x_t - 0.95 \cdot x_{t-1}$ that is applied to the speech signal samples x_t. Obviously, the two methods have many similarities. Differences between PLP and MFCC lie in the filter-banks, the equal-loudness pre-emphasis, the intensity-to-loudness conversion and in the application of LP. Each of them is discussed in the following.

As we consider the discrepancy between the progression of

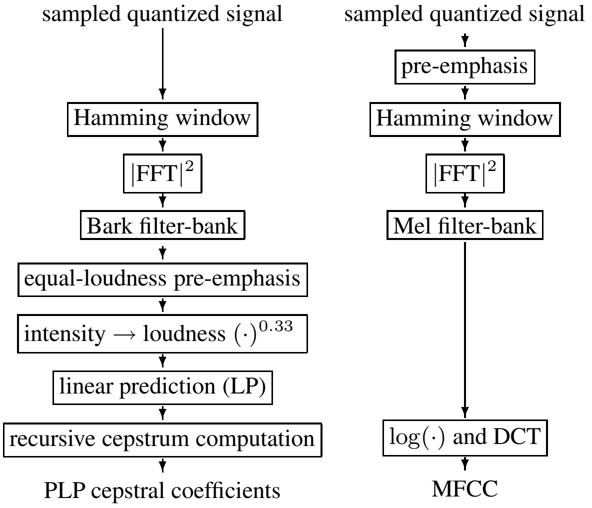

Fig. 1: *The computation steps of PLP (left) and MFCC (right).*

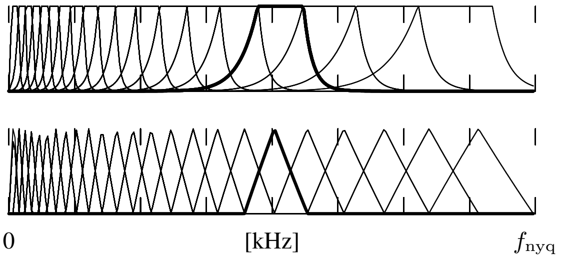

Fig. 2: *Bark (top) and Mel filter-bank for a power spectrum with 257 coefficients (not normalized for display purposes).*

the Bark- and the Mel-scale to be negligible in practice (see e. g. [8, p. 34]), the most prominent difference between Bark and Mel filter-bank is the shape and, presumably more important, the number and the width of the filters. It can be seen in Fig. 2 that the Bark filter-bank consists of 19 asymmetrically-shaped filters while the Mel filter-bank contains typically 24–40 triangular filters which have a 50%-overlap (e. g. [8, p. 317]).

Hermansky [2] introduced an equal-loudness pre-emphasis of the power spectrum to take into account the frequency sensitivity of human hearing. Each power spectrum coefficient $P(\omega)$ is multiplied with a weight $E_1(f) = \frac{(f^2+1.44\cdot 10^6)f^4}{(f^2+1.6\cdot 10^5)^2(f^2+9.61\cdot 10^6)}$ that depends on its frequency in Hertz $f = \frac{\omega}{2\pi} f_{\mathrm{sample}}$. For signals with a Nyquist frequency f_{nyq} that is greater than 5 kHz, an alternative weighting function $E_2(f)$ is defined in [2, Eq. 7'] which gives a better approximation of the psycho-physical findings as it also represents the decrease in sensitivity in the higher frequency range. Both weighting functions are shown in Fig. 3. As it is shown in Fig. 3 the equal-loudness weighting function E_1 used in PLP is quite similar to the pre-emphasis that is applied to the speech signal in the MFCC computation. The effect of the pre-emphasis has to be discussed in the context of the successive application of LP. The relation between a discrete input power spectrum $P(\omega)$ and the corresponding LP model power spectrum $\hat{P}(\omega)$ is given by [9]:

$$\frac{1}{M}\sum_{m=1}^{M}\frac{P(\omega_m)}{\hat{P}(\omega_m)} = 1 \qquad (1)$$

Thus, $\hat{P}(\omega)$ is a smooth fit to $P(\omega)$; the smoothness is enforced by the fact that $\hat{P}(\omega)$ is an all-pole spectrum of low order. It

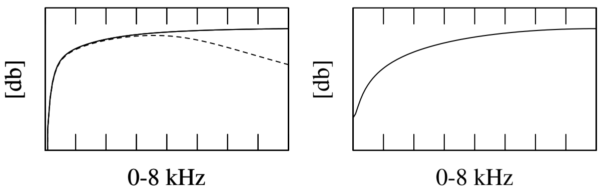

Fig. 3: *The equal-loudness weighting function E_1 in PLP (left, solid) is quite similar to a conventional pre-emphasis applied to the signal (right). The equal-loudness weighting function E_2 is shown on the left, dashed.*

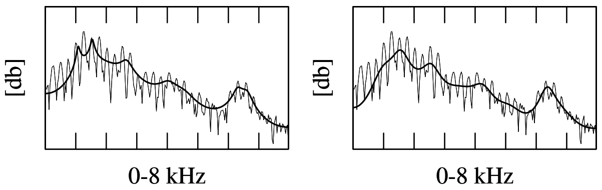

Fig. 4: *Model spectrum of a linear prediction of order 20 (left). Right: with $(.)^{0.33}$ compression prior to LP (scaled). In each plot, the original power spectrum is also shown as a thin line.*

can be seen from Eq. 1 that large values of $P(\omega)$ have more influence on the overall fit than small values. As in speech signals high frequencies have much lower magnitudes than low frequencies, it is beneficial to equalize the power spectrum before applying LP. The application of a pre-emphasis increases high frequencies and thus makes the spectral fit of LP more uniform over the frequency range.

Fig. 4 (left) shows that the LP model approximates the spectral envelope of $P(\omega)$. This is due to Eq. 1. However, together with the all-pole model assumption, this leads to sharp peaks in $\hat{P}(\omega)$. The intensity-to-loudness conversion, which raises the power spectrum coefficients to the power of 0.33, decreases the dynamic variability and flattens the peaks of $P(\omega)$. The resulting model spectrum is smoother with less pronounced peaks as it is demonstrated in Fig. 4 (right). The intensity-to-loudness conversion is therefore to be seen as a tuning of the spectral envelope approximation.

3. Modifications of PLP

The discussion of the different processing steps of PLP and the comparison with the MFCC computation motivate several modifications of PLP. In particular, we measure experimentally which of the two filter-banks (Bark or Mel filter-bank) leads to better performance and propose an improved filter-bank for PLP and MFCC. Further experiments investigate the pre-emphasis and the computation of the LP.

3.1. Filter-bank

There seems to be no intuitive reason why the Bark filter-bank should be optimal for PLP (see Sec. 2) and other filter-bank configurations should be evaluated as well. A promising candidate is the popular Mel filter-bank which has already successfully been employed within PLP by others [7]. Fig. 5 illustrates that the application of the Mel filter-bank is equivalent to three consecutive processing steps: (i) Mel frequency warping of the power spectrum; (ii) convolution with a triangle (i. e. smoothing); and (iii) down-sampling. The number of filters resembles the precision of the sampling in step (iii). In order to minimize

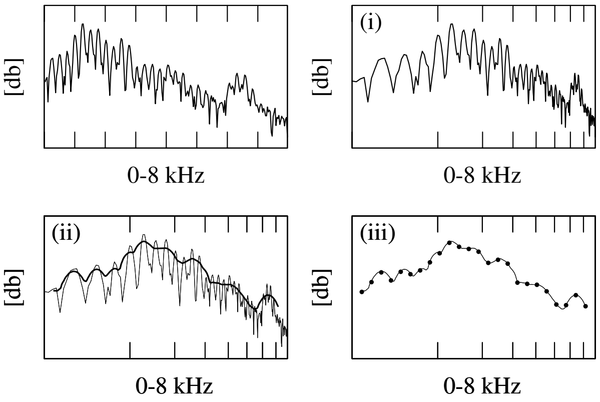

Fig. 5: *The steps implicitly performed when the standard Mel filter-bank (24 filters of width* 226.8 *Mel) is applied: original power spectrum; warping; smoothing; sampling.*

the loss in information, the number of filters should be as high as possible. In the conventional definition of the Mel filter-bank, the triangular filters have an overlap of 50% (e. g. [8, p. 317]); i. e. increasing the number of filters reduces at the same time their band-width. The width of the filters, however, defines the characteristics of the smoothing step and should be determined independently. Therefore we propose an alternative definition of a Mel filter-bank: the filter band-width is fixed to 226.8 Mel, which is equivalent to the band-width in a conventional Mel filter-bank with 24 filters. The number of filters is chosen independently; here we take 257 filters which is the number of input spectral coefficients. Consequently, these filters have an overlap which is much larger than 50%.

3.2. Pre-emphasis

We have shown in Sec. 2 that there is a high degree of similarity between the equal-loudness weighting performed in PLP and the pre-emphasis that is applied to the speech signal in the conventional MFCC computation. The latter is widely used and can be applied in a very simple manner before the short-time spectral analysis. Furthermore, it is not clear if E_1 performs better than E_2 or if both give approximately the same performance. Therefore we have to measure experimentally which one of the three different methods to perform pre-emphasis (E_1, E_2 or to the signal) leads to the best results.

3.3. LP input

The center frequencies of the first and last Bark filter have a distance of ≈ 1 Bark to the boundaries of the frequency range $[0; f_{\mathrm{nyq}}]$. According to Hermansky, there should be two additional filters with center frequencies 0 and f_{nyq}, but they cannot be computed because the filters would reach too far into undefined frequency ranges [2]. As a substitute, the first and last filter-bank output are duplicated in conventional PLP (after the equal-loudness pre-emphasis). We take a different view here and discard the duplication of the boundary values in order to avoid an over-emphasis[1]. The effective frequency range of the LP is then determined by the center frequencies f_1, f_k of the first and last filter as $[f_1; f_k]$. Note that it is possible to move the center-frequencies of the filters to 0 and f_{nyq} by mirroring and continuing the power spectrum.

4. Data sets and baseline system

Training and evaluation were done on two different corpora: ChildIt and HUB4. ChildIt has been recorded at ITC-irst with a close-talk microphone and consists of Italian texts read by children. It comprises 8 h of speech for training and 2.5 h for evaluation. The language model was estimated on articles taken from Italian newspapers. For the experiments on HUB4, we used the BN-E data released by the LDC in 1997 and 1998 as training data. These corpora contain a total of about 143 h of speech. The 1998 Hub4 evaluation data consists of 3 h of speech *(Eval98)*. For selected experiments, we report the HUB4 results w. r. t. the focus conditions *(F-conditions)* marked in Eval98:

F0:	baseline planned broadcast speech, clean background.
F1:	spontaneous broadcast speech, clean background.
F2:	speech over telephone, clean background.
F3:	speech with background music.
F4:	speech with degraded acoustics (noise, other speech).
F5:	planned, non-native speech, clean background.
FX:	all other conditions that cannot be classified into F0-F5.

Language models were trained on approx. 132 million words of broadcast news transcripts distributed by LDC and on the transcripts of the BN-E training data. The acoustic front-end of the ITC-irst speech recognition system applies cluster-based mean and variance normalization to the first 13 cepstral coefficients and combines them with their first and second order time derivatives into a 39-dimensional feature vector. In all experiments, we use $\mathrm{Mel}(\omega) = 1125 \cdot \ln(1 + \omega/700)$. No manual labeling of clusters is used in training and recognition. The acoustic models are state-tied, cross-word, gender-independent, bandwidth-independent triphone HMMs. The MFCC baseline system has 701 tied states and about 11000 Gaussians for ChildIt; for HUB4 it has 9079 tied states and about 146000 Gaussians. All other systems in this work have a similar number of parameters.

5. Experimental results

The word error rates (WER) for the MFCC and PLP baseline systems are shown in the first two columns of Tab. 1. While PLP performs 4.6% rel. worse than MFCC on the ChildIt task which has been recorded in a clean acoustic environment, we obtain about equal results for both feature types on HUB4 which is acoustically less homogeneous. Systems (A)–(D) in Tab. 1 correspond to different variants of PLP. In system (A) the Bark filter-bank is replaced by a conventional Mel filter-bank with 24 filters. This leads to a noticeable improvement on both data sets, however, results on ChildIt are still worse than for MFCC. Next, systems (A), (B) and (C) compare different ways to perform the pre-emphasis. System (A) and (B) apply the weighting functions E_1 and E_2, respectively. System (C) applies a pre-emphasis directly to the speech signal like in MFCC, this is denoted by the letter S in Tab. 1. While there is no measurable difference between E_1 and E_2, the pre-emphasis applied to the signal gives a better performance for ChildIt and

Table 1: *WER in % for MFCC and different variants of PLP.*

system	*MFCC*	*PLP*	*(A)*	*(B)*	*(C)*	*(D)*	*RPLP*
filter-bank	Mel	Bark	Mel	Mel	Mel	Mel	Mel
pre-emphasis	S	E_1	E_1	E_2	S	E_1	S
duplicate values	-	√	√	√	√	-	-
$(\cdot)^{0.33}$ and LP	-	√	√	√	√	√	√
ChildIt	15.2	15.9	15.6	15.6	15.3	15.2	**14.8**
HUB4	20.5	20.6	20.1	20.1	20.2	20.1	**19.9**

[1] A compromise between Hermansky's procedure and our approach would be to apply DCT-II [8, p. 228] when computing the auto-regression coefficients. However, in our experiments we found this to be sub-optimal.

a slightly higher WER for HUB4. By discarding the duplication of the filter-bank values at the boundaries we derive system (D) and system *revised PLP* (RPLP) from the systems (A) and (C), respectively. While in both cases discarding the duplication leads to improvements in WER, the best results are obtained for system RPLP, i. e. the combination of a pre-emphasis applied to the speech signal together with the discarded boundary values. An explanation for the fact that system (C) performs worse than (A) on HUB4 could be that the harmful effect of the duplication step is more pronounced in combination with the pre-emphasis than with the weighting functions E_1 and E_2, where the weight for the first filter output is almost zero (see Fig. 3). The revised PLP is equivalent to the baseline MFCC feature computation with an intensity-to-loudness conversion and a LP step. System RPLP performs better than both baseline systems on both tasks. More precisely, the rel. improvements in WER are 6.9% and 3.4% over conventional PLP for ChildIt and HUB4, respectively, and 2.6% and 2.9% over MFCC for ChildIt and HUB4, respectively. The differences between systems RPLP and MFCC and between RPLP and PLP are significant according to the matched pairs sentence-segment word error (MAPSSWE) test. For the HUB4 Eval98 test set the respective p-values are 0.002 and < 0.001.

Once we have found in system RPLP a suitable configuration for PLP, we compare the conventional Mel filter-bank (24 filters) with the proposed Mel filter-bank that consists of 257 filters as described in Sec. 3.1. The corresponding results are shown in Tab. 2. For the HUB4 task, results are also given for the second step, i. e. after the application of MLLR adaptation with two regression classes. For MLLR the first-step result of each system is taken as a supervision. From Tab. 2 it can be seen that the new filter-bank with 257 filters leads to additional improvements on the HUB4 task both for MFCC and PLP. On the ChildIt corpus, however, there is only an improvement for the MFCC features, while for RPLP the new filter-bank results in a slight increase in WER[2]. We found it encouraging to note that the improvements obtained by revised PLP and the proposed filter-bank stay the same or become even more prominent after MLLR adaptation (see Tab. 2). The baseline systems are compared with the revised PLP for the different focus conditions of the Eval98 test set in Tab. 3. It can be seen that RPLP together with the proposed filter-bank performs better than both MFCC and PLP for nearly all marked conditions. Remember that non-native speech (F5), the only condition where it performs worse than PLP, corresponds to less than 1% of the test data. The overall rel. improvements in WER w. r. t. the MFCC and PLP baseline systems are 3.9% and 4.4%, respectively. After MLLR, the revised PLP with 257 filters performs in both cases 4.8% better than the MFCC and PLP baseline systems.

Table 2: *WER [%] for different filter-bank configurations.*

system	ChildIt	HUB4	
		1st step	2nd step
MFCC, 24 filters	15.2	20.5	18.7
MFCC, 257 filters	**14.8**	**20.2**	**18.4**
PLP, 19 filters	15.9	20.6	18.7
RPLP, 24 filters	**14.8**	19.9	18.2
RPLP, 257 filters	15.0	**19.7**	**17.8**

Table 3: *WER [%] on HUB4 w. r. t. the focus conditions.*

F-condition	all	F0	F1	F2	F3	F4	F5	FX
proportion	100.0	30.7	19.3	3.4	4.3	28.2	0.7	13.5
MFCC	20.5	12.8	20.3	31.5	24.0	20.7	26.0	33.9
PLP	20.6	12.9	19.8	31.0	23.3	21.2	20.4	34.5
RPLP, 257 f.	19.7	12.1	19.3	30.1	21.1	19.9	25.1	33.5

6. Conclusion and future work

We have shown that the PLP computation can be improved noticeably. Our revised setup for PLP applies a pre-emphasis to the signal, and employs a Mel filter-bank with a large number of filters (e. g. 257) and a band-width around 230 Mel. Equal-loudness weighting and duplication of the boundary values of the filter-bank are discarded. This setup simplifies the computation of PLP. Compared to conventional PLP, we reduce the WER by 5.7% rel. for the ChildIt corpus and by 4.4% rel. for HUB4. The improvements remain after MLLR adaptation. MFCC benefits from the proposed filter-bank as well.

Due to the higher sampling resolution of the proposed filter-bank, we expect it to be beneficial when used together with Vocal Tract Length Normalization (VTLN) and other normalization approaches like the one described in [10]. Therefore we are investigating the combination of the revised features with an adaptive training procedure [10].

7. Acknowledgment

This work was partially financed by the European Commission under the projects TC-STAR (ref. number FP6-506738) and HUMAINE (ref. number FP6-507422) and by the German Federal Ministry of Education and Research within the SmartWeb project (grant no. 01 IMD 01 F).

8. References

[1] S. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. ASSP*, vol. 28, no. 4, pp. 357–366, 1980.

[2] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal Acoust. Soc. Amer.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[3] P. Woodland, M. Gales, and D. Pye, "Improving environmental robustness in large vocabulary speech recognition," in *Proc. ICASSP*, vol. 1, 1996, pp. 65–68.

[4] E. Schukat-Talamazzini, *Automatische Spracherkennung – Grundlagen, statistische Modelle und effiziente Algorithmen.* Braunschweig: Vieweg, 1995.

[5] M. Hunt, "Spectral signal processing for ASR," in *IEEE ASRU Workshop*, 1999.

[6] J. Mason and Y. Gu, "Perceptually-based features in ASR," in *IEEE Colloquium on Speech Processing*, 1988, pp. 7/1–7/4.

[7] P. Woodland, M. Gales, D. Pye, and S. Young, "The development of the 1996 HTK broadcast news transcription system," in *DARPA Speech Recognition Workshop*, 1997.

[8] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing – A Guide to Theory, Algorithm, and System Development.* Upper Saddle River: Prentice Hall, 2001.

[9] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[10] G. Stemmer, F. Brugnara, and D. Giuliani, "Adaptive training using simple target models," in *Proc. ICASSP*, vol. 1, 2005, pp. 997–1000.

[2] Note that by widening the frequency range covered by the center-frequencies of the filters to $[0; f_{\mathrm{nyq}}]$ as indicated in Sec. 3.3, we can reach a WER of 14.6% with the proposed filter-bank. However, we believe that such optimizations are strongly corpus-dependent.