IDIAP RESEARCH REPORT

COMPARISON AND COMBINATION OF FEATURES IN A HYBRID HMM/MLP and a HMM/GMM SPEECH RECOGNITION SYSTEM

Pere Pujol, Susagna Pol, Climent Nadeu, Astrid Hagen, Hervé Bourlard

IDIAP-RR 03-48

SEPTEMBER 2003

Dalle Molle Institute
for Perceptual Artificial
Intelligence • P.O.Box 592 •
Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

COMPARISON AND COMBINATION OF FEATURES IN A HYBRID HMM/MLP AND A HMM/GMM SPEECH RECOGNITION SYSTEM[#]

Pere Pujol [*], *Susagna Pol, Climent Nadeu*[*],
Talp Research Center, Universitat Politècnica de Catalunya (UPC)
Campus Nord, Edifici D5, C/ J. Girona 1-3, 08034 Barcelona, Spain. Tel. (+34) 93 401 6438
Fax: (+34) 93 401 6447 E-mail: [**pujol**, **climent]@talp.upc.es**

Astrid Hagen
INESC-ID, Spoken Language Systems Lab (L2F). Rua Alves Redol, 9-2, 1000-029
Lisbon, Portugal. Tel. (+351) 21 310 03 13 E-mail: **astrid.hagen@weenie.inesc.pt**

Hervé Bourlard
IDIAP. Rue du Simplon, 4, C.P. 592, CH-1920 Martigny, Switzerland.
Tel. (+41) 277 217 720 Fax. (+41) 27 721 77 12 E-mail: **bourlard@idiap.ch**

ABSTRACT

Recently, the advantages of the spectral parameters obtained by frequency filtering (FF) of the logarithmic filter-bank energies (logFBEs) have been reported. These parameters, which are frequency derivatives of the logFBEs, lie in the frequency domain, and have shown good recognition performance with respect to the conventional MFCCs for HMM systems. In this paper, the FF features are first compared with the MFCCs and the Rasta-PLP features using both a hybrid HMM/MLP and a usual HMM/GMM recognition system, for both clean and noisy speech. Taking advantage of the ability of the hybrid system to deal with correlated features, the inclusion of both the frequency second-derivatives and the raw logFBEs as additional features is proposed and tested. Moreover, the robustness of these features in noisy conditions is enhanced by combining the FF technique with the Rasta temporal filtering approach. Finally, a study of the FF features in the framework of multi-stream processing is presented. The best recognition results

for both clean and noisy speech are obtained from the multi-stream combination of the J-Rasta-PLP features and the FF features.

1. INTRODUCTION

The conventional mel-frequency cepstral coefficients (MFCC), which are obtained from a discrete cosine transformation of the log filter-bank energies (logFBEs), are widely used in speech recognition. In spite of their popularity, they have some disadvantages: first, they do not lie in the frequency domain, and, secondly, as most current HMMs use Gaussian distributions with diagonal covariance matrices, these HMMs cannot benefit from cepstral weighting. In order to avoid these drawbacks, Nadeu et al. [1] and, more recently, Paliwal [2] have proposed the use of a new kind of logFBE-based features which result from a simple linear transformation, frequency filtering (FF), that is able to quasi-decorrelate the logFBEs. These new FF features, which are briefly presented in Section 2, are actual derivatives of the logFBEs along frequency, and have generally shown a better recognition performance than the MFCCs, especially for noisy speech [3].

The FF features had been tested so far in an HMM system based on Gaussian mixture modeling (GMM). In this paper, they are tested in the framework of a hybrid recognition system that uses a multi-layer perceptron (MLP) to compute the emission probabilities. In addition, recognition experiments are also performed with a HMM/GMM recognition system, in order to compare results from both recognizers. In Section 3, both the systems and the used digit database are presented, including an investigation of the spectral characteristics of the added real noises. In Section 4, experimental tests with the digit utterances for both clean and noisy speech are reported. In them, alternative FF features resulting from the concatenation of the usual first-derivative features with the second-derivative ones and the (non-filtered) logFBEs are employed.

Additionally, the behavior of the FF technique in combination with other known techniques that enhance the robustness of the recognition system is studied in this paper. In Section 5, the FF

technique is combined with the Rasta approach, a temporal high-pass filtering applied to the logFBEs, and its experimental results are compared with the ones from the Rasta-PLP (perceptual linear prediction) features [4,5].

The power of the multi-stream technique [6] has been employed with the HMM/MLP system to obtain further improved results with a combination of the FF features with the J-Rasta-PLP features using a product rule. As reported in Section 6, for clean speech, a WER improvement of 18% with respect to the conventional HMM/GMM system with MFCCs is obtained with that multi-stream approach, and higher improvements are observed in almost all noise conditions.

Figure 1

2. THE FREQUENCY FILTERING TECHNIQUE AND SOME EXTENSIONS

In order to use logFBEs as features of an HMM speech recognition system based on Gaussian mixture densities, it can be useful to decorrelate them, since one of the assumptions in most current HMMs is the diagonality of the feature covariance matrices. The frequency filtering technique is a linear transformation that is able to almost decorrelate the logFBEs by convolving them with the impulse response of a FIR filter [1,3]. As it is shown in [1,3], this operation can also be seen as a cepstral liftering.

In the FF technique used in this work, the sequence of logFBEs of a given frame, i.e.

$$S_1,S_2,\ldots,S_k,\ldots,S_Q \qquad (1)$$

where Q is the number of bands, is convolved with the impulse response of either the 1st or the 2nd order FIR filter presented in [1] to obtain a new sequence of filtered parameters, which will be referred to FF features. The impulse responses and the transfer functions of the 1st order (FF1) and 2nd order (FF2) filters are, respectively:

$$h_1(k) = \{+1,-1\} \qquad H_1(z) = 1 - z^{-1}$$

(2,3)

$$h_2(k) = \{+1,0,-1\} \;\; H_2(z) = z - z^{-1} \qquad (4,5)$$

FF2 has generally shown a better speech recognition performance than FF1 using HMM/GMM systems, so that FF2 is the usual version for frequency filtering [1,3]; but in this work it will be shown that FF1 may be competitive for hybrid systems and, under determined noisy conditions, for GMM-HMM systems.

Concerning the application of the filtering idea to the sequence of logFBEs, there are several alternatives. We can use only the differences of the computed logFBEs, but this is not a true convolution, since the transients are discarded. To include the transients, we may consider that the logFBE sequence is extended with zeroes. In this work, as it is suggested in [1,3], when the FF1 filtering is used, a zero is appended at the beginning of the logFBEs sequence in order to compute the first element of the filtered sequence. Similarly, when the 2nd order FIR filter is used, zeros are added at both extremes of the logFBE sequence. The value zero is reasonable since if we extend the mel-scaled filter-bank with one more (triangular) filter at each side, we will collect a small fraction of the signal energy (which it is not useful for speech recognition, as shown by the fact that the usual filter-banks do not include these end filters).

Thus, the resulting sequence after filtering with FF1 is

$$(F_1^1, F_2^1, \mathrm{K}\ , F_Q^1) = (S_1 - 0, S_2 - S_1, ..., S_Q - S_{Q-1})$$

(6)

and, the resulting sequence after filtering with FF2 is

$$(F_1^2, F_2^2, \mathrm{K}\ , F_Q^2) = (S_2 - 0, S_3 - S_1, ..., 0 - S_{Q-1})$$

(7)

In this way, the first element of the sequences in (6-7) actually is an absolute energy measure, as well as the last element in the case of the FF2 filtering. As shown in previous papers, this choice yields good recognition performance, at least for clean speech. Actually, it is a way to include the frame energy in the feature set. However, as occurs with the energy feature in usual ASR systems, it may happen that, by excluding those absolute energies, an improvement in terms of recognition performance is observed. For example, when the speech signal is corrupted by low-pass noise, the first FF feature may be removed to improve speech recognition performance.

The other elements in (6-7) are differences of logFBEs, i.e. coarse measures of the slope of the spectral envelope. In 1982, D. Klatt showed that a phonetic distance based on the spectral slopes near the peaks correlates very well with perceptual data, unlike distances based on other speech characteristics such as FBEs [7].

Herewith, the frequency-filtered logFBEs (FF features) resulting from FF1 and FF2 filtering are referred to **FF1** and **FF2** features, respectively. Extensions of the frequency filtering technique, consisting in the addition of other frequency features to the FF feature vector, as the raw logFBEs and a new kind of features, which will be called twice-frequency-filtered logFBEs, were experimented. As it is done with the logFBEs to obtain the FF features, the frequency sequence of FF features are convolved with the corresponding filter (FF1 or FF2) response to calculate the twice-frequency-filtered parameters. Thus, the resulting sequences are, respectively:

$$(F_1^{1'}, F_2^{1'}, \mathrm{K}\ , F_Q^{1'}) = (F_1^1 - 0, F_2^1 - F_1^1, ..., F_Q^1 - F_{Q-1}^1) \quad (8)$$

$$(F_1^{2'}, F_2^{2'}, \mathrm{K}\ , F_Q^{2'}) = (F_2^2 - 0, F_3^2 - F_1^2, ..., 0 - F_{Q-1}^2) \quad (9)$$

This operation is depicted in Figure 2. Features resulting from applying two times the FF1 or FF2 filters are referred to **FF1'** and **FF2'** respectively. Moreover, we refer to the logarithmically compressed FBEs as **FBE**.

Figure 2

The above feature sets may be concatenated in several ways to compose the final parameter vector. Figure 3 shows the various feature sets that are considered in this paper.

Figure 3

3. EXPERIMENTAL SETUP

3.1. Recognition systems: HMM/GMM and hybrid HMM/MLP

The principle aim of an HMM/ANN (Artificial Neural Network) hybrid system is to combine the efficient discriminative learning capabilities of neural networks and the superior time warping and decoding techniques associated with the HMM approach. The ANN is trained to estimate HMM emission probabilities which are then used by a Viterbi decoder for recognition [8]. The ANN used in this paper is a multi-layer perceptron (MLP). It is known that the hybrid systems have some advantages over traditional HMM systems. Using an MLP, no assumptions about the statistical distributions of the input features are necessary. Due to its classification procedure, an MLP has the ability to decorrelate the input features. Moreover, while in a classical HMM-based system the parameters are trained according to a likelihood criterion, an MLP also penalizes the incorrect classes.

To choose the size of the MLP hidden layers, several tests were carried out for each different number of input features, since there is no specific rule relating both parameters. For the case of one set of features, 1750 neurons are chosen, and this number is increased to 2500 in the case of

using two sets of features, and to 3000 in the case of three sets. It is designed to train monophone posterior probabilities $P(q_l \mid x_j)$, where x_j is the input vector and q_l is a monophone, so that, the output layer contains a number of neurons equal to the number of phones in the data base (27 monophones). In order to provide the MLP with contextual information, 9 consecutive frames of data are given as input.

The other HMM-based recognizer chosen for the experiments is the HMM/GMM (Gaussian Mixture Model). In this approach, the observation probabilities are modelled as a weighted sum of Gaussian probability densities. The system uses 80 triphones. Each speech unit is represented by an HMM with 5 states (the first and the last ones are non-emitting) and 10 mixture components per state.

Finally, a decoder implementing the Viterbi algorithm finds the state sequence having the highest probability of generating the observation sequence in both systems. The word transition probability was optimised to give the best results for clean speech. In the case of the hybrid system, the use of posterior probabilities or scaled likelihoods to directly replace the state probability densities which are normally modelled by GMMs was optimized for each test. The referred scaled likelihoods are obtained dividing the posterior probabilities at the output of the MLP by the prior probabilities of the phone P(q_l).

3.2. Database and noise characteristics

Both recognizers were trained and tested with the Numbers95 database [9] (naturally spoken digits over the telephone line). All trainings were carried out using 3233 utterances of clean speech data and 357 for the cross-validation test. A set of 1206 utterances (4670 words) was used for testing. In the case of tests with noisy speech, those clean speech files were contaminated with additive noise. Two different real-environmental noises were used: car noise provided by Daimler Chrysler, and factory noise from Noisex92 [10]. Tests were carried out at different SNR levels

(18 dB, 12 dB, 6 dB and 0 dB). The SNR was computed excluding the silence portions of the waveforms. The car noise can be considered as a stationary noise whereas the factory noise is less stationary, has impulsive regions in time (hammer blows), and shows energy peaks in the spectral formants region, so that it deteriorates more speech recognition performance than the car noise. In order to simulate noisy speech recorded over the telephone line, the noise signals were filtered with a band-pass filter with cut-off frequencies 216 Hz and 3770 Hz, which is similar to the one used in telephony in.

In order to evaluate the impact of the noises on the recognition systems, it may be useful to know the distribution of the noise over the different frequency bands used in the feature extraction process. In our experiments with FF features, 12 mel-scaled bands are employed. The distribution of the spectral energy for clean speech (over the whole Numbers95 database) along these bands can be seen in Figure 4. The signal-to-noise ratio (SNR) in each of these 12 bands for car- and factory-corrupted speech, when the SNR level is 18 dB, is shown in Figures 5 and 6, respectively. The SNR has been computed by averaging the speech spectral energies and the noise spectral energies over all the speech frames of the test utterances (note that clean speech signals and noise signals are available separately). Figures 5 and 6 also show the SNR of each band for the case of band-pass-filtered noisy speech.

Figure 4.

As can be observed in Figure 5, the first band is very corrupted by the car noise since this type of noise has most of its energy concentrated in the low-frequency part of the spectrum. The other bands have a higher and nearly constant SNR. The figure also shows how the use of the band-pass filtering strongly increases the SNR of this corrupted part of the spectrum. However, the removal of the high-frequency part of the spectrum with band-pass filtering does not noticeably affect the SNR measure of the last FBEs.

Figure 5

As shown in Figure 6, factory noise not only affects the low frequencies as occurs with the car noise, but it also harms the highest and middle part of the spectrum. This characteristic, together with its non-stationary nature, makes it very harmful for speech recognition, as it will be seen in the experiments. Again, removing the high-frequency part of the spectrum with band-pass filtering does not noticeably affect the SNR measure of the last FBEs, but the first two bands are largely favoured by that signal filtering.

Figure 6

4. RECOGNITION EXPERIMENTS USING THE FF TECHNIQUE

In this section, results obtained using the FF features with either a HMM/GMM or a hybrid HMM/MLP system are reported. Furthermore, experiments using MFCC's were carried out in order to compare both techniques. Dynamic features corresponding to the usual first and second time derivatives are appended to the initial static vector in all the experiments.

4.1. Experiments using the HMM/GMM system

4.1.1. Results with clean speech

For these experiments, speech was pre-emphasized with a first order FIR filter with a zero at z=0.95, windowed with a Hamming window of 30 ms and shifted every 12.5 ms. Then, 12 mel-scale filter-bank energies were computed for each frame and logarithmically compressed. The word (digit) error rates (WER) yielded by the feature sets resulting from FF1 and FF2 frequency filters are compared with that from MFCCs in Table 1. The best results obtained from the

concatenated FF feature sets from Figure 3 are also shown in the table. For the results with clean speech presented in the forthcoming tables, the 95% confidence interval is about ± 0.7.

Table 1.

Recognition results from 12 filtered FBEs obtained with the second-order FIR filter (FF2) are better, although not statiscally significant, than results from the conventional MFCCs, which include 13 cepstral coefficients (obtained from 26 bands) and the frame energy. However, none of the tested concatenations could obtain good results in comparison with the FF2 features, which are the usual FF features [3]. This is probably due to the strong correlation between the feature vectors included in each concatenation and within the elements of the FBE vector.

4.1.2. Results with noisy speech

In the experiments performed with speech corrupted by car noise (Table 2), features obtained by means of FF2 filtering (FF2) gave the best results, as it was the case in clean speech. It can be noticed that the performance of the FF1 filtering (FF1) is harmed by car noise since its first parameter equals the first FBE, which is the FBE most corrupted by this noise. On the other hand, the behaviour of the FF1 technique in factory noise conditions (Table 3) is better than the FF2 one. In experiments with the concatenated FF features, it was observed that, like for clean speech, they do not seem a good alternative for noisy speech using a HMM/GMM system.

Table 2.

Table 3.

The total results presented in Tables 2 and 3 were computed by averaging the results from the various SNRs for each type of noise. For these results, the 95% confidence intervals are about ± 0.4 for the car noise and ± 0.7 for the factory noise.

4.2. Experiments using the hybrid HMM/MLP system

4.2.1. Results with clean speech

To get the FF features for the input to the hybrid HMM/MLP system, the input speech signal was pre-emphasized and windowed with a Hamming window of 25 ms shifted every 12.5 ms. Then, 12 mel-scale filter-bank energies were computed for each frame and logarithmically compressed to obtain the logFBEs. For the FF1 features, the preemphasis zero is at z=0.97, and for the FF2 ones, it is at z=0.95.

A comparison of the word error rates of various FF features and that of the MFCCs (computed as pointed out in Section 4.1.1.) is presented in Table 4.

Table 4.

All the features obtained with the FF technique gave better performance than MFCCs. Moreover, a slight improvement is obtained from the new extensions of the FF technique (Figure 3) as more sets of features are added (FBE+FF2 and FBE+FF2+FF2'). Actually, only the differences corresponding to these two kinds of concatenated features are statistically significant. Thus, the 3-set concatenation got the lowest error rate. It is interesting to notice that, for HMM with Gaussian mixture densities, FF2 usually yields higher recognition scores than FF1 for clean speech [1-3], but in our tests with the HMM/MLP system, the results from FF2 and FF1 have been identical.

4.2.2. Results with noisy speech

For the experiments with the HMM/MLP system in noisy speech, the band-pass filter mentioned in Section 3.2 was used to remove the noise at both sides of the spectrum. It consisted of applying the filter bank from 216 Hz to 3770 Hz, instead of applying it to all the frequency range. This band-pass filtering didn't affect the clean speech recognition performance due to the fact that the Numbers95 database was collected through a telephone line, so the removed frequency

components are almost absent from the clean speech database. In addition, as both noises have strong components at the lowest frequencies of the spectrum, they are partially removed with this filtering. Actually, in the case of the simulated car environment, the noise energy is largely removed.

In Table 5, the word error rates obtained with the FF feature sets for car noise experiments are presented in comparison with the results from MFCCs. As it occurred with clean speech, the best performance was achieved by FBE+FF2+FF2', for all the tested SNRs. In addition, although differences are not statistically significant, MFCCs results are again surpassed by those from FF2 and FF1 features for all SNRs. Total results shown in Tables 5 and 6 have the same confidence intervals as Tables 2 and 3.

Table 5.

In factory noise conditions, the best frequency filtering features are based on the FF1 filtering, as it can be seen in Table 6. This is probably due to the fact that the FF1 features do not contain any absolute energy measure of the highest part of the spectrum, which is the part with the lowest SNR, as shown in Figure 6. In particular, the FF1+FF1' concatenation got the lowest error rates in comparison with the other tested methods. It seems to occur that the FBE features are so severely corrupted by this noise that, when they are added to the concatenated feature vector, the results worsen. In fact, in some experiments with the FF2 filtering and factory noise, using either FBE, FF2 or FF2' features separately, the FBE features got the worst scores. It is also interesting to note that the best results in those experiments were obtained by FF2' features.

Table 6.

In summary, for the hybrid system, the best one-set version of FF features is inour tests, the first derivative obtained from FF1. These FF1 features are able to outperform the FF2 features in the factory environment case, and work almost equally well for both clean speech and speech

corrupted by car noise. Furthermore, for all the tested environments, the results from the FF1 features can be improved with the concatenation of various sets of frequency-filtered features.

4.3. Comparison of results from the two recognition systems

For clean speech, the usual FF features, i.e. FF2, outperform in our tests the widely used MFCC features in both systems, HMM/GMM and HMM/MLP. Unlike the MFCCs, the FF features lie in the frequency domain, and their computational cost is even lower than that of MFCCs, since the DCT is substituted by a simple set of Q subtractions.

In addition, although the new approach based on the concatenation of features obtains in our tests the best score with the hybrid HMM/MLP recognizer, that score is still lower than the one obtained with the HMM/GMM system and employing the basic FF2 features. As we have observed, when the concatenations are tested with the HMM/GMM system, the performance gets worse. A summary of the results for clean speech obtained in this section are presented in Figure 7.

Figure 7.

A comparison between the two systems for noisy speech is depicted in Figures 8 and 9. It can be observed that the behaviour of the various features for car noise (Figure 8) is similar to their behaviour for clean speech. This can be due to the fact that the car noise is stationary and basically harms only the first band.

Figure 8.

The experiments with car noise show that the HMM/MLP system, using the concatenation of 3 sets of features FBE+FF2+FF2' clearly improves, for low SNR, the best results using the HMM/GMM system, which correspond to the FF2 features. However, a triple number of spectral features are needed and we must notice that the HMM/MLP system includes a band-pass filtering

which attenuates the noise severely. Nonetheless, we should note that when the band-pass filtering was employed in the HMM/GMM recognizer, the results did not improve, being that the reason why the use of band-pass filtering was avoided in that system.

In the case of factory noise (see Figure 9), both systems achieve the best results with frequency-filtered features using the FF1 filter. Again, the HMM/GMM system yields a lower WER than the hybrid system for both SNR=12 and 18 dB; but for lower SNRs, the best result correspons to the hybrid system using the concatenation of FF1 and FF1' features.

Figure 9.

For both systems, an improvement of the FF technique (either FF1 or FF2, or sometimes both) over the MFCC parameters can be observed in all experiments, except when the when the HMM/GMM system is tested with factory noise and SNR=12 or 18 dB (see Table 3). Actually, it appears that the FF features are especially harmed by the factory noise used in our tests, and that the FF2 features are more affected by this noise than the FF1 features. As the SNR of the factory-corrupted speech at high frequencies is rather low, as it is shown in Figure 6, this noise can harm the performance of the FF2 filtering since its last parameter is an absolute energy measure of the highest part of the spectrum. The use of FF1 filtering avoids this problem, and this can be the reason of its better performance in all the reported factory noise experiments.

5. FF FEATURES WITH RASTA FILTERING

The RASTA (RelAtive SpecTrA) approach [5] is based on a band-pass time-filtering applied to a log-spectral representation of the speech, such as the log filter-bank energies. As it is explained in [5], the high-pass component of the Rasta filter can alleviate the effect of the convolutional noise introduced in the channel while the low-pass filtering is expected to help in smoothing out some fast frame-to-frame spectral changes present in the short-term spectrum of the speech signal.

Our aim is to evaluate the effect of Rasta filtering on the features studied in this paper, i.e. FF features and concatenated FF features, since experiments developed in [3] have shown that the recognition performance of frequency-filtered logFBEs can be improved with specific temporal filtering. Only FF2 filtering is used.

5.1. Methodology

The methodology that was followed to compute the Rasta-FF features is the following:

- Extraction of 12 log filter-bank energies from each speech frame.
- Filtering of the FBE's with the 4th-order Rasta filter (from [5]), which has the following transfer function:

$$H(z) = 0.1\frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4}(1 - 0.98\,z^{-1})} \quad (10)$$

- Application of the FF technique to the modified FBEs resulting in a new kind of features which will be called Rasta-FF features.
- Concatenation of the Rasta-FF features (as it was done with the concatenation of FF features).

5.2. Experimental results

The resulting features were tested with an HMM/MLP recognizer and an HMM/GMM one, with clean and noisy speech, using the same test sets as in Section 4. In order to put the results in a wider perspective, some recognition tests were carried out with the Rasta-PLP features [5]. These features were obtained applying the Rasta technique to the Perceptual Linear Prediction (PLP) features [4], which try to model a perceptually-motivated spectrum by an all-pole model function

using the autocorrelation linear prediction technique. In our experiments, 12 PLP-cepstral coefficients were obtained from 17 critical bands, plus the energy.

5.2.1. HMM/GMM system

In Table 7, a comparison between FF2, Rasta-FF2 and Rasta-PLP features is shown. It can be observed that the Rasta filtering applied to the FF2 features (Rasta-FF2) approximately maintains the recognition performance of the HMM/GMM system for clean speech and high-SNR car noise, and gets a WER improvement in the rest of tests with noisy speech. The tests with Rasta-PLP features yield a significantly worse WER than the Rasta-FF2 features for both clean and car-noise-corrupted speech.

Although, for the factory environment, the Rasta-FF2 features get an improvement of 14% (18dB) and 27% (6dB), they are not able to perform as well as Rasta-PLP features in that noise condition. Concerning this observation, we recall the comments from the last paragraph of Section 4.

Table 7.

In Table 7, the interval for a 95% of confidence is about ± 0.4 for car noise results, whereas in the case of factory corrupted speech, the confidence intervals are ± 1.4 and ± 0.9 for SNR= 6 and SNR=18 respectively.

5.2.2. HMM/MLP system

According to the results shown in Table 8, in the HMM/MLP system, when the FF2 frequency filtering is combined with the Rasta technique, noticeable improvements are achieved in most of the tested environments, especially for noisy speech. The combination of the concatenated features with Rasta shows the lowest WER in all the tested conditions, even outperforming the

Rasta-PLP features in factory noise tests. In this table, the confidence intervals are the same as in table 7. On the other hand, some experiments performed with FF1-based features such as FF1 or FF1+FF1', showed that the combination of that frequency filter with Rasta technique harmed the recognition performance.

Table 8.

From our tests, we can conclude that both the hybrid system and the HMM/GMM system usually get better recognition results when the Rasta filtering is applied to the FF features. This suggests that both filtering techniques, one of which is working over time, the other over frequency, may cancel out different noise components in the signal. Moreover, even though the performance of the HMM/MLP system is noticeably improved by including the temporal Rasta filtering, the results cannot reach yet those from the HMM/GMM system, except for the high-SNR car noise condition.

6. MULTI-STREAM APPROACH WITH FF AND J-RASTA-PLP FEATURES

Some of the state-of-the-art ASR systems employ multi-stream processing [6,11]. In multi-stream processing, several data streams are processed in parallel before their information is recombined at a later point. Combination of the different streams can be carried out either before or after acoustic modeling, i.e. on the feature level or on the probability level.

In the case of feature combination, specific conditions are learned by the recognizer which works on the combined stream and which, thus, will have difficulties to generalize in the case where one stream offers well-known data but the other stream is in a harmful way different from the data represented in the training data. As feature combination of two or more streams, moreover,

usually leads to a rather large feature vector, which thus also demands larger models and more training data, probability combination might be preferred if training data is sparse. Moreover, the expected increase in robustness might be achieved more easily, if the different feature representations can each be fully exploited by one specific recognizer instead of forcing one recognizer to work on all representations.

In probability combination, the respective stream-recognizers learn the specific good data of that stream and are not disturbed by unrecognized data in the other feature streams. Thus, probability combination seems to offer increased robustness as compared to feature combination.

For this reason we decided to investigate the multi-stream approach which employs probability combination to recombine the different input streams.

The multi-stream technique can be easily implemented in an HMM/MLP system [6]. Given a phone q_l, the probabilities $P(q_l|x_j)$ obtained by the MLP from every independent stream of data x_j are combined via the product rule to get the probabilities to input to the Viterbi decoder. In our case, each stream x_j corresponds to a different parameterization of the signal frame. The formula used is:

$$P(q_l|\mathbf{x}) \approx \frac{\prod_{j=1}^{R} P(q_l|x_j)}{P^{R-1}(q_l)} \qquad (11)$$

where $\mathbf{x}$ is the set of R streams $\mathbf{x}=(x_1,...,x_j,...,x_R)$, $P(q_l|\mathbf{x})$ is the result of combining the probabilities $P(q_l|x_j)$ from all the R different classifiers (one for each stream), and $P(q_l)$ is the prior probability of the phone q_l calculated as the relative frequencies of phone q_l in the training set. This product rule for posterior combination was derived from the well-known product rules for likelihoods using the Bayes' rule.

The posterior probabilities theoretically need to be divided by the class prior probabilities to obtain (scaled) likelihoods for Viterbi decoding. As it has been found during experimental evaluation [6], this division does not always lead to improved performance, depending on the respective features, database and other conditions. For this reason, we evaluated in preliminary experiments for which system the division by priors was necessary. In our experiments with the FF-based and J-Rasta-PLP recognizers we found that the division by priors is harmful in most of the environments. It was therefore decided for the least loss in performance, that is when outputs of the two recognizers are not divided by the priors.

In our experiments, we combine feature streams using two different approaches. One is the product rule formula for posterior probabilities (11), and the other consists in the simple multiplication of posterior probabilities, which was employed in [12]. This rule assumes independence of the posterior probabilities of one class given the data from different streams. As we will see it is shown in this section, the product rule (11) as derived from the likelihood case [6] leads to more robust results than the independence assumption for posteriors.

6.1. Experimental results

Two different sets of streams were employed in the framework of FF features. First, tests were carried out combining feature streams which both stem from frequency filtering but with different filters (FF1 and FF2). As can be seen in Table 9, the product rule outperforms the simple multiplication of probabilities. Additionally, the product rule achieved lower WERs than the single-stream systems where each stream was used by itself (except for factory noise with SNR=6). Unfortunately, these results from FF multi-stream processing can still not surpass the best recognition performance shown in the previous sections. In this Table, 95% confidence ranges are the same as in Table 7.

Table 9.

Next, we investigated the use of more diverse features, aiming at finding features that are synergistically combined by the multi-stream approach. In this way, we have found experimentally that the combination of FF features with J-Rasta-PLP features is rather powerful, whereas the combination with either MFCCs or Rasta-PLP features did not lead to any improvement of the system (actually, as all of them are computed from logFBEs). The J-Rasta approach is a variation of the so-called Rasta technique, which consists in the application of Rasta filtering in an alternative spectral domain which is linear-like for small spectral values and logarithmic-like for large values [13]. It should be noticed that this technique uses a parameter called J whose optimal value depends on the SNR. In our experiments, the value of the parameter J was set to 10^{-6}.

Tables 10, 11 and 12.

Tables 10-12 show the WERs from the multi-stream system employing the FF features in one stream and the J-Rasta-PLP features in the other stream. The 95% confidence intervals are about ± 0.4 for the total results with car noise and ± 0.7 for the factory noise.

For clean speech and car noise the product rule resulted in a statistically significant lower WER than the baseline HMM/MLP system with either FF or J-Rasta-PLP features. For car noise the tendency is similar to that in clean speech. A relative improvement of at least 25% (SNR = 6dB) over the FF2 features were obtained. The simple multiplication of probabilities also resulted in a statistically significant reduction in WER as compared to each of the feature used by itself, though the difference is not as high as when the product rule was used.

For factory noise, the J-Rasta-PLP features turned out to be significantly more robust than any of the previously tested features. A further improvement was obtained when the J-Rasta-PLP and FF feature stream were used jointly in the multi-stream approach employing the product rule (except

for SNR=0dB). It is interesting to notice that this improvement could be achieved although the FF feature stream used alone produced a high WER in this kind of noise. In this case, combination by simple multiplication is not able to outperform the single-stream system employing the J-Rasta features.

To sum up, in clean speech, the multi-stream approach obtained a WER improvement of 18% as compared to the conventional HMM/GMM system using MFCC features. Higher improvements were observed in almost all noise conditions. These results could be even further when the FF2 feature stream was substituted by the FBE+FF2+FF2' stream, although this also provides a higher computational cost. We can conclude that the application of J-Rasta filtering to the PLP features produces a type of feature (J-Rasta-PLP) which is complementary to the FF features when both are combined in a multi-stream system. In this system, product rule usually outperformed the simple multiplication in our tests carried out here.

7. CONCLUSIONS

The main goal of our work was to study the FF features in the framework of a hybrid HMM/MLP recognition system. In addition, FF features were also experimented with a HMM/GMM recognition system, in order to compare both recognizers. First of all, we observed that the speech recognition performance of the FF features was better than the performance of the widely used MFCCs in most tested cases, in spite of the lower computational cost associated to the FF features. Using the hybrid system, the results of the FF features were improved by concatenating them with other frequency features. With this extended set of features, the hybrid system was able to (slightly) outperform the HMM/GMM system for noisy speech and low SNRs.

The combination of the Rasta technique with the frequency filtering technique enhanced the robustness of the FF features in noisy conditions for both recognition systems. Nevertheless, the best recognition performance for both clean and noisy speech was observed using the multi-

stream HMM/MLP approach with a product rule, and combining the FF features with the J-Rasta-PLP features. For clean speech, a WER improvement of 18% with respect to the conventional HMM/GMM system with MFCCs was obtained with that multi-stream approach, and higher improvements were observed in almost all noise conditions.

8. REFERENCES

[1] C. Nadeu, J. Hernando and M. Gorricho (1995). “On the Decorrelation of Filter-Bank Energies for Speech Recognition”, Proc. Eurospeech, pp. 1381-1384.

[2] K. K. Paliwal (1999). “Decorrelated and Liftered Filter-Bank Energies for Robust Speech Recognition”, Proc. Eurospeech, pp. 85-88.

[3] C. Nadeu, D. Macho and J. Hernando (2001). “Frequency and Time Filtering of Filter-Bank Energies for Robust HMM Speech Recognition”, Speech Communication (Special Issue on Noise Robust ASR), Vol. 34, pp. 93-114.

[4] H.Hermansky (1990). ”Perceptual Linear Predictive (PLP) Analysis of Speech”, J. Acoust. Soc. Am., 87(4), pp.1738-1752.

[5] H. Hermansky, N. Morgan (1994). "Rasta Processing of Speech", IEEE Trans. on Speech and Audio Proc. Vol.2, No.4, pp.578-589.

[6] A. Hagen, H. (2001). “Robust Speech Recognition Based on Multi-Stream Processing”, (PhD Thesis), École Polytechnique Fédérale de Lausanne, Switzerland.

[7] D.H. Klatt (1982). “Prediction of Perceived Phonetic Distance from Critical Band Spectra: a First Step”, Proc. ICASSP, pp. 1278-1281.

[8] H. Bourlard and N. Morgan (1994). “Connectionist Speech Recognition. A Hybrid Approach”, Kluwer Academic Publishers, Massachussets, 1994.

[9] R.A. Cole, M. Noel, T. Lander and T.Durham (1995). “New Telephone Speech Corpora at CSLU”, Proc. Eurospeech, Vol. 1, pp. 821-824.

[10] A. Varga, H.J.M. Steeneken, M. Tomlison and D. Jones (1992). “The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition”, Documentation included in the NOISEX-92 CD-ROM Set.

[11] D. Ellis and M.J. Reyes Gomez (2001). “Investigations into Tandem Acoustic Modeling for the Aurora Task”, Proc. Eurospeech, Vol. 1, pp. 189-192.

[12] K. Kirchhoff and J. Bilmes (2000). “Combination and Joint Training of Acoustic Classifiers for Speech Recognition”, Proc. ISCA ITRW Workshop on Automatic Speech Recognition (ASR2000), pp. 17-23.

[13] J. Koehler, N. Morgan, H. Hermansky, H.G. Guenter and G. Tong (1994). “Integrating Rasta-PLP into Speech Recognition”, IEEE Trans. on Signal Processing, Vol. 1, pp. 421-424.

LIST OF FIGURES AND TABLES

FIGURES AND TABLES

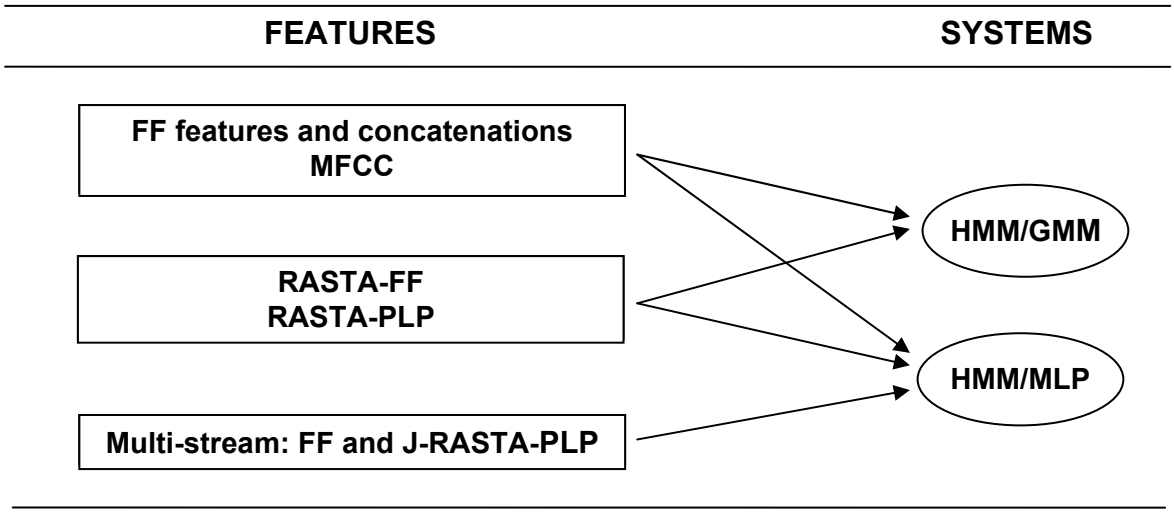

Figure 1: Block diagram summarizing the various experiments carried out in this work. All experiments are carried out with clean speech, and speech corrupted by car and factory noises at different SNRs (0, 6, 12, 18 dB).

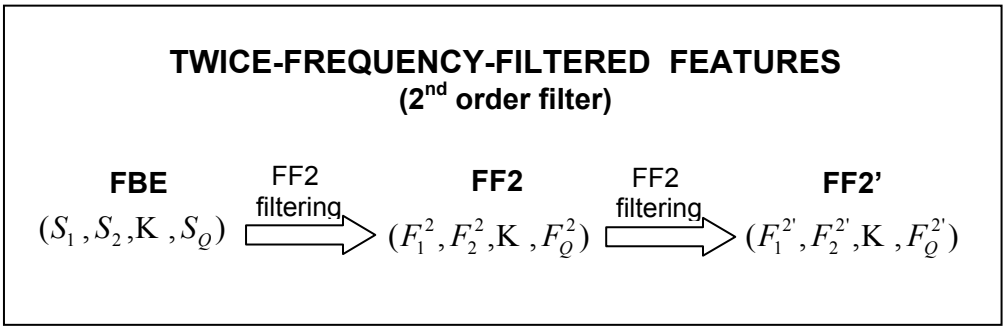

Figure 2: Twice-frequency-filtered features for the case of FF2 filtering (**FF2'**).

FF FEATURE SETS
(2nd order filter)

1 SET

FBE: $(S_1, S_2, \mathrm{K}, S_Q)$

FF2: $(F_1^2, F_2^2, \mathrm{K}, F_Q^2)$

FF2': $(F_1^{2'}, F_2^{2'}, \mathrm{K}, F_Q^{2'})$

2 SET

FBE + FF2: $(S_1, \ldots, S_Q, F_1^2, \ldots, F_Q^2)$

FF2 +FF2': $(F_1^2, \ldots, F_Q^2, F_1^{2'}, \ldots, F_Q^{2'})$

3 SET

FBE + FF2 +FF2': $(S_1, \ldots, S_Q, F_1^2, \ldots, F_Q^2, F_1^{2'}, \ldots, F_Q^{2'})$

Figure 3: The various FF feature sets employed in this work, for the case of FF2 filtering.

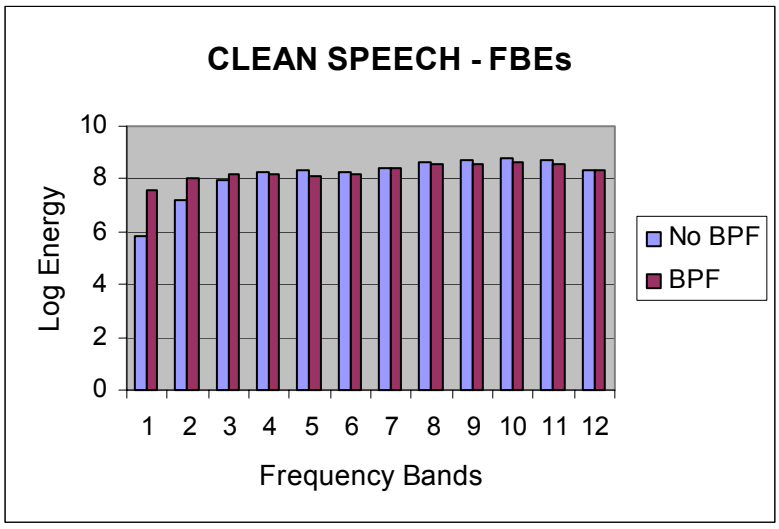

Figure 4: Distribution of the energy of the clean speech signal along the frequency bands, with and without a band-pass filtering (BPF).

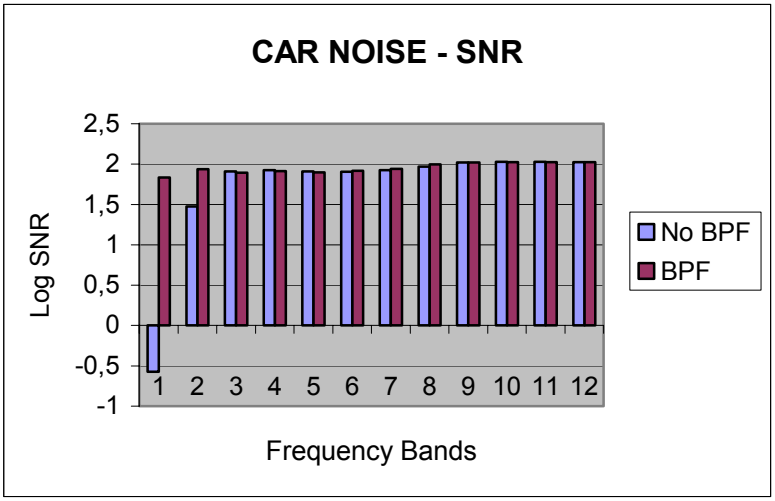

Figure 5: Distribution of the mean SNR along the frequency bands in files corrupted by car noise, with and without band-pass filtering (BPF)

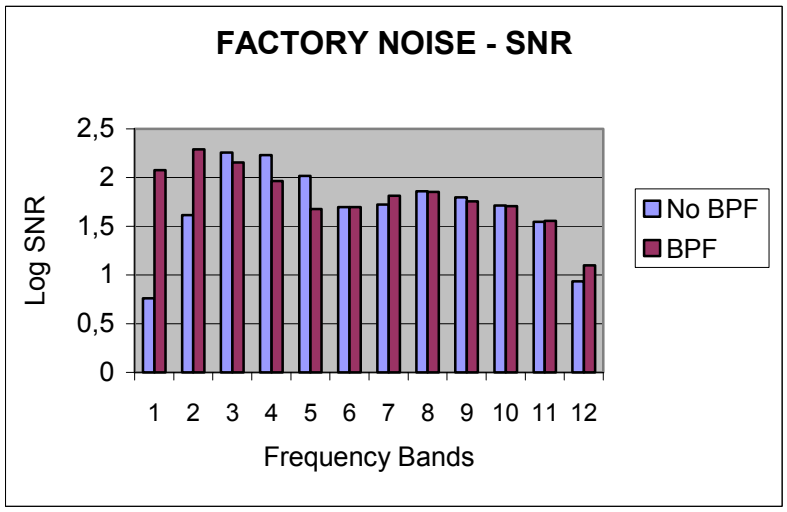

Figure 6: Distribution of the mean SNR along the frequency bands in files corrupted by factory noise, with and without band-pass filtering (BPF)

WER	Clean Speech
MFCC	6.7
FF1	7.4
FF2	6.2
FF2+FF2’	7.4
FBE+FF2+FF2’	8.0

Table 1: Results, given in terms of word error rate (WER), for MFCCs and FF features with clean speech using the HMM/GMM system.

WER	SNR = 0	SNR = 6	SNR =12	SNR = 18	Total
MFCC	10.7	8.5	7.5	7.0	8.4
FF1	14.1	10.6	8.5	8.1	10.3
FF2	8.6	7.1	6.6	6.5	7.2

Table 2: Test results for MFCC, FF1, and FF2, under **car noise** conditions using the HMM/GMM system

WER	SNR = 0	SNR = 6	SNR =12	SNR = 18	Total
MFCC	78.8	35.6	15.3	9.4	34.8
FF1	70.7	33.8	16.5	10.4	32.8
FF2	86.8	42.0	18.5	10.7	39.6

Table 3: Test results for MFCC, FF1, and FF2 under **factory noise** conditions using the HMM/GMM system

WER	Clean Speech
MFCC	7.3
FF1	6.8
FF2	6.8
FBE + FF2	6.5
FBE+FF2 +FF2’	6.3

Table 4: Test results for MFCCs and FF features with clean speech using the HMM/MLP hybrid system

WER	SNR = 0	SNR = 6	SNR =12	SNR = 18	Total
MFCC	8.7	8.1	8.1	8.2	8.3
FF1	8.1	7.5	7.6	7.5	7.7
FF2	8.1	7.7	7.8	7.7	7.8
FBE+FF2+FF2’	7.6	6.8	6.7	6.7	6.9

Table 5: Test results for MFCC, FF1, FF2, and FBE+FF2+FF2’ under **car noise** conditions using the hybrid system

WER	SNR = 0	SNR = 6	SNR =12	SNR = 18	Total
MFCC	71.1	37.9	21.4	13.6	36
FF1	63.0	33.9	19.3	12.9	32.2
FF2	70.2	37.5	20.5	13.7	35.5
FF1+FF1’	61.1	32.5	18.4	12.2	31

Table 6: Test results for MFCC, FF1, FF2, and FF1+FF1’ under **factory noise** conditions using the hybrid system

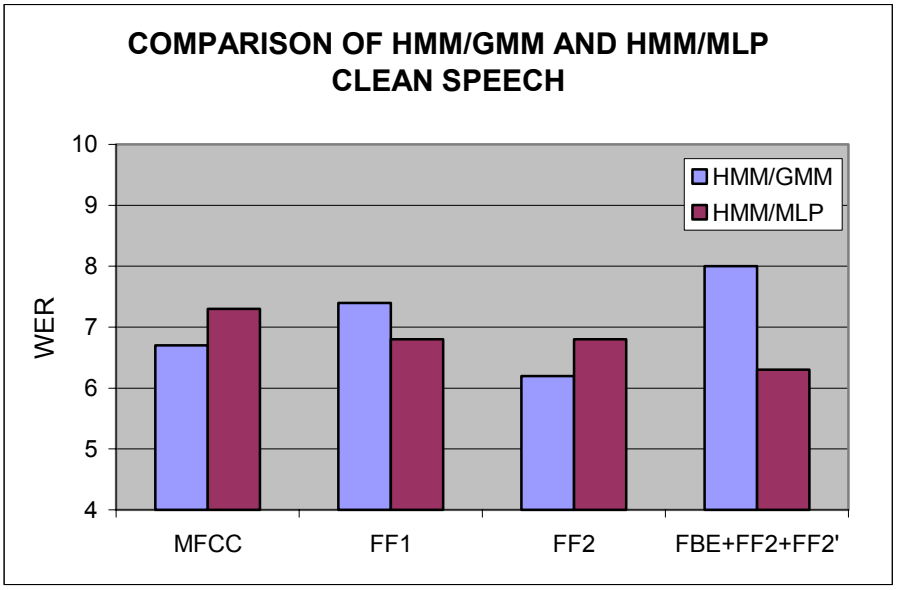

Figure 7: Comparison of HMM/GMM and HMM/MLP recognizers with clean speech

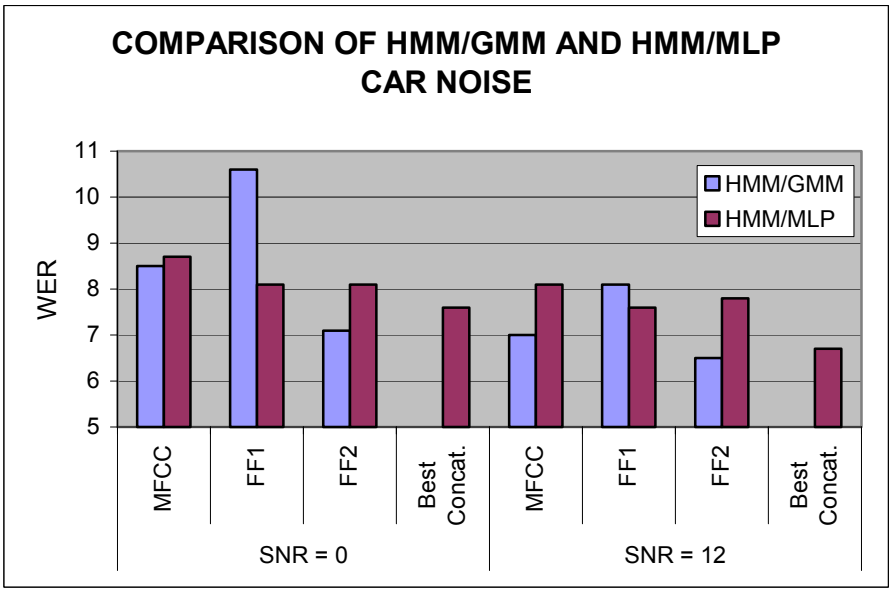

Figure 8: Comparison of HMM/GMM and HMM/MLP recognizers with speech corrupted by **car noise**. The results from the HMM/GMM system with concatenated features are omitted

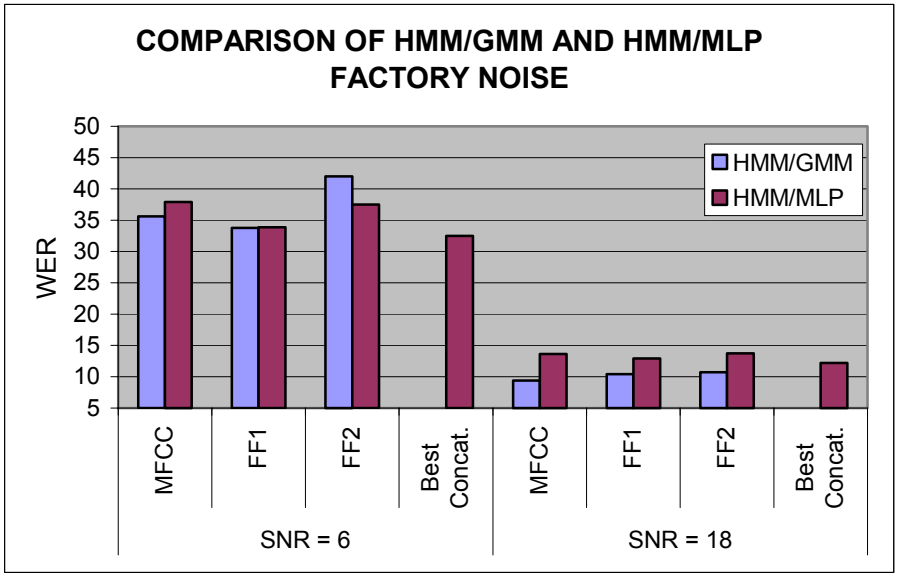

Figure 9: Comparison of HMM/GMM and HMM/MLP recognizers with speech corrupted by **factory noise**. The results from the HMM/GMM system with concatenated features are omitted

WER	Clean	Car noise		Factory noise	
		SNR=6	SNR=18	SNR=6	SNR=18
FF2	6.2	7.1	6.4	42.0	10.7
Rasta-FF2	6.2	6.7	6.5	30.7	9.2
Rasta-PLP	6.7	8.8	7.7	28.1	8.7

Table 7: Test results for Rasta-PLP and Rasta-FF features with clean and noisy speech using the HMM/GMM recognizer

WER	Clean	Car noise		Factory noise	
		SNR=6	SNR=18	SNR=6	SNR=18
FF2	6.8	7.7	7.7	37.5	13.7
Rasta-FF2	6.6	7.0	7.0	33.1	10.9
Rasta-FBE+FF2+FF2'	6.3	6.5	6.4	31.2	10.8
Rasta-PLP	6.5	9.1	6.4	31.6	11.5

Table 8: Test results for Rasta-PLP and Rasta-FF features with clean and noisy speech using the HMM/MLP recognizer

WER	Clean	Car noise		Factory noise	
		SNR=6	SNR=18	SNR=6	SNR=18
FF1	6.8	7.7	7.7	33.9	12.9
FF2	6.8	7.5	7.5	37.5	13.7

Multiplication	6.6	7.6	6.9	34.8	12.8
Product rule	6.1	6.6	6.5	35.7	11.3

Table 9: Multi-stream combination of FF2 and FF1 features with clean and noisy speech

WER	Clean Speech
FF2	6.8
J-Rasta-PLP	6.8
Multiplication	6.1
Product rule	5.5

Table 10: Multi-stream combination of FF2 and J-Rasta-PLP features with clean speech

WER	SNR = 0	SNR = 6	SNR =12	SNR = 18	Total
FF2	8.1	7.7	7.8	7.7	7.8
J-Rasta-PLP	9.6	9.1	8.7	9.0	9.1
Multiplication	6.5	6.4	6.3	5.6	6.2
Product rule	6.0	5.8	5.6	5.6	5.7

Table 11: Multi-stream combination of FF2 and J-Rasta-PLP features, in **car noise** conditions

WER	SNR = 0	SNR = 6	SNR =12	SNR = 18	Total
FF1	63.0	33.9	19.3	12.9	32.3
J-Rasta-PLP	41.8	21.1	12.8	9.7	21.3
Multiplication	50.0	24.2	13.4	9.6	24.3
Product rule	43.4	20.6	11.3	8.0	20.8

Table 12: Multi-stream combination of FF1 and J-Rasta-PLP features, in **factory noise** conditions