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 mel-frequency cepstral coefficients (MFCC) for HMM systems. In this paper, the FF features are compared with the MFCCs and the Rasta-PLP features in the framework of a hybrid HMM/MLP 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 the second frequency derivatives and the raw logFBEs as additional features is proposed. Furthermore, in order to enhance the robustness of these features in noisy conditions, they are combined with the Rasta temporal filtering approach. Finally, a study of the FF in the framework of multistream processing is presented. From the experimental tests, it appears that the new spectral parameters and the tested combinations yield an enhanced recognition performance.

1. INTRODUCTION

It is well known that mel-frequency cepstral coefficients (MFCC), even though they show good performance in speech recognition, also have some disadvantages. First, they do not lie in the frequency domain. 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, actually are 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 densities. In this paper, they are tested in the framework of a hybrid HMM/MLP recognition system [4], which is presented in Section 3. In Section 4, experimental tests with both clean and noisy speech are reported by concatenating the usual first-derivative features with the second-derivative ones and with the (non-filtered) logFBEs.

In the second part of the article, the behavior of the FF technique in combination with other known techniques that enhance the robustness of the recognition system are studied. In Section 5, the FF technique is combined with Rasta, which is a temporal high pass filtering applied to the logFBEs, and its experimental results are compared with the ones from the PLP-Rasta features [5]. Furthermore, the behavior of FF in combination with the multi-stream technique is studied in Section 6. 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.

2. FREQUENCY FILTERING (FF) TECHNIQUE AND SOME EXTENSIONS

In order to use logFBEs as features of an HMM speech recognition system based on Gaussian mixture densities, they must be de-correlated, since one of the main assumptions in such a recognizer is the diagonality of the feature covariance matrices. The frequency filtering technique [1] is a linear transformation that is able to almost decorrelate the logFBEs by convolving them with the impulse response of an FIR filter, an operation which also can be seen as a cepstral liftering.

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

\[ S = (S(1), S(2), ..., S(Q)) \]

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. The impulse responses and the transfer functions of the 1st order (FF1) and 2nd order (FF2) filters are:

\[ h_{FF1}(k) = \{+1, -1\} \]  \[ H_{FF1}(z) = 1 - z^{-1} \]  \[ (2,3) \]

\[ h_{FF2}(k) = \{+1, 0, -1\} \]  \[ H_{FF2}(z) = z - z^{-1} \]  \[ (4,5) \]

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. On the other hand, when the 2nd order FIR filter is used, zeros at both extremes of the logFBE
sequence are added. The resulting vectors after the filtering operation are,
\[ F_{FF1}(k) = \{ S(1) - 0, S(2) - S(1), ..., S(Q) - S(Q - 1) \} \] (6)
\[ F_{FF2}(k) = \{ S(2) - 0, S(3) - S(1), ..., 0 - S(Q - 1) \} \] (7)
The first element of the resulting vectors actually is an energy measure, as well as the last coefficient in the case of the FF2 filtering. In this paper, features resulting from FF2 and FF1 filtering are called df and ddf respectively.

Extensions of the frequency filtering technique, consisting in the addition of other frequency features, as the raw logFBEs (called fb) or twice frequency filtered FBEs, were experimented. We will call df and ddf the set of features resulting from applying two times the FF2 and FF1 filters respectively, to the logFBEs. These new features are concatenated to compose the final parameter vector (See Figure 1). It was expected that this additional information renders the system more robust to noise. Dynamic features computed by first and second time derivatives are appended to the initial static vector in all the experiments.

In order to provide the MLP with contextual information, 9 consecutive frames of data are given as input.

Finally a decoder implementing the Viterbi algorithm finds the state sequence having the highest probability of generating the observation sequence in both systems.

4. RECOGNITION EXPERIMENTS OF THE FF TECHNIQUE WITH A HYBRID SYSTEM

In this section, the most important results obtained using the FF features with a hybrid HMM/MLP system are commented. The MLP was trained and tested with the Numbers95 database (naturally spoken digits over the telephone line). All trainings were carried out using 3233 utterances (4670 words) of clean speech data and 357 for the cross-validation test. A set of 1206 utterances was used for testing. In the case of noise tests, these files were contaminated with additive noise. Two different kinds of noises were used (car noise from a database provided by Daimler Chrysler and factory noise from Noisex92) at different SNR levels (18 dB, 12 dB, 6 dB and 0 dB). The word transition probability was optimised to give the best results in clean speech and the ‘divide-by-priors’ parameter was also optimized for each test. This parameter, when is set to 1, causes the division of the output probabilities of the MLP by the prior probabilities of the phone to get the scaled likelihoods needed by the Viterbi decoder.

In all the experiments with noisy speech shown in this article, the filter bank was applied between 216 KHz and 3770 KHz, instead of spreading the filter bank over all the frequency range, in order to remove the noise in both sides of the spectrum. This band pass filtering doesn’t affect the recognition with clean speech but improves the recognition with noise since Numbers95 database is a telephone speech database and actually, the frequencies cut are not correctly transmitted by the telephone line. In addition, as both noises are highly concentrated in the lowest frequencies of the spectrum, they are partially removed. In the case of car environment, noise energy is removed almost entirely with this filtering.

4.1. Results with clean speech

To get the FF features, first the input speech signal was pre-emphasized, then filtered with a Hamming window of 25 ms taken every 12.5 ms and 12 MEL-scale filter-bank energies computed for each. Then, these energies were logarithmically compressed to obtain the logFBEs. The features resulting from FF2 filtering worked better using a pre-emphasis filter with a zero at z=0.95 and FF1 worked better with zero at z=0.97.

A comparison of the word error rates (WERs) of various FF releases and optimized MFCCs (13 cepstra, including a log energy obtained from 26 FBEs) is presented in Table 1.

<table>
<thead>
<tr>
<th>WER</th>
<th>MFCC</th>
<th>df1</th>
<th>df</th>
<th>fb + df</th>
<th>fb + df + ddf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>7.3</td>
<td>6.8</td>
<td>6.8</td>
<td>6.5</td>
<td>6.3</td>
</tr>
</tbody>
</table>

Table 1: Results for MFCCs and FF features with clean speech.
densities, FF1 usually yields lower recognition scores than FF2 for clean speech, but in our tests with an HMM/MLP system the result has been identical.

4.1.1. Results with noisy speech

In Table 2, the most relevant word error rates obtained with the FF feature sets for car noise experiments are presented in comparison with the results of MFCCs. As it happened with clean speech, the performance of the 3-set features (i.e. fb+df+ddf) is better than the other features for all SNRs tested. It may be due to the fact that the extra information added to the raw FF is not so corrupted by this kind of noise. In addition, MFCCs results are again surpassed by df and df features.

In factory noise conditions, the best frequency filtering features are based on the FF1 filtering, as it can be seen in Table 3, meanwhile the performance of the df features is very similar to that of the MFCCs.

<table>
<thead>
<tr>
<th>FEATURES</th>
<th>SNR = 0</th>
<th>SNR = 6</th>
<th>SNR = 12</th>
<th>SNR = 18</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>8.7</td>
<td>8.1</td>
<td>8.1</td>
<td>8.2</td>
</tr>
<tr>
<td>dff</td>
<td>8.1</td>
<td>7.5</td>
<td>7.6</td>
<td>7.5</td>
</tr>
<tr>
<td>df</td>
<td>8.1</td>
<td>7.7</td>
<td>7.8</td>
<td>7.7</td>
</tr>
<tr>
<td>fb+df+ddf</td>
<td>7.6</td>
<td>6.8</td>
<td>6.7</td>
<td>6.7</td>
</tr>
<tr>
<td>Table 2: Test results for MFCC, dff, df, and fb+df+ddf under car noise conditions</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In this case, the concatenation of once and twice filtered logFBEs achieved the best recognition rates. It can be concluded that fb features are so corrupted by this noise that if they are added to the concatenated feature vector (as in fb+df or fb+df+ddf), results get worse. In fact, in some experiments using either fb, df or ddf features separately, fb features got the worst scores with factory noise. It is also remarkable that the best results were obtained by ddf features.

5. RASTA FILTERING WITH FF FEATURES

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. Our aim is to evaluate the effect of the Rasta filtering on the features studied in this paper, i.e. FF features and concatenated FF features, since the ability of RASTA processing to deal with various kinds of noise has been demonstrated, and experiments developed in [1] have, moreover, shown that frequency filtered logFBEs can be improved with specific temporal filtering.

The filter used for the experiments was extracted from [5]:

\[ H(z) = 0.1 \frac{2 + z^{-1} - z^{-3} - 2z^{-2}}{z^{-1}(1 - 0.98 z^{-1})} \]  

5.1. Experimental results

To get the Rasta-FF features, the Rasta filtering was applied to 12 logFBEs. The modified logFBE’s were then filtered using the FF technique. The resulting features were tested in an HMM/MLP recognizer 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 [5] features. These features were obtained applying the Rasta technique to the Perceptual Linear Prediction (PLP) features [6], which try to model a perceptually motivated spectrum by an all-pole model function using the autocorrelation LP technique.

Two completely different behaviors can be observed when Rasta technique is combined with FF1 and FF2 based features. In the case of FF2 frequency filtering, the application of Rasta technique improves greatly the system in most of the environments tested. In Table 4, a slight improvement can be seen when the RASTA filters were used together with the df and fb+df+ddf features in a clean environment, although the most notable improvement was obtained with noisy speech. In a car noise environment, relative improvements of about 9% can be appreciated with df features and the same happens to a stronger extent with factory noise, achieving a 13% for SNR=18.

Moreover, Rasta-fb+df+ddf features show a superiority over the rest of the features in almost all the cases.

6. MULTI-STREAM

In order to improve the robustness of the FF features in front of noise, multi-stream processing technique was experimented. The probabilities obtained from each stream of data with the MLP, \( P(q_k | x) \), were combined via the product rule from [7] to get the probabilities to input to the Viterbi decoder. The formula used is:

\[ P(q_k | x) \approx \prod_{j=1}^{R} \frac{P(q_k | x_j)}{P_{1:R}(q_k)} \]  

where \( x=(x_1,\ldots,x_{2R}) \) consisting of R independent acoustic vectors (one for each stream), \( P(q_k | x) \) is the result of combining the probabilities for the phoneme \( q_k \) of all
the different classifiers (one for each stream), and \( P(q_k) \) is the prior probability of the phoneme \( q_k \) calculated by the relative frequencies of each phoneme in the training set.

Our study was focused in the search of some features, which combined with FF, could obtain an improvement of both of the systems involved.

6.1. Experimental results

Whereas the combination of FF features with MFCCs or Rasta-PLP features didn't contribute to an improvement of the system (they are all based on the logFBEs), a very positive effect was obtained when the FF features were combined with J-Rasta-PLP features. The J-Rasta approach is a variation of the so-called Rasta technique. It 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 [5]. It should be noticed that, there is a parameter called J and its optimal value varies for different SNRs. In our experiments, it was set to \( 10^{-6} \).

In the tests with clean speech, Table 5, the combination of df and J-Rasta-PLP features achieved a WER of 6.1\%, decreasing in 10\% the WER of the combined features. Results obtained with car noise, Table 5, show similar improvements to those with clean speech but in the case of factory noise, the FF WER is so high that the combination can not improve the J-Rasta-PLP results.

Furthermore, a new and effective way to combine the posterior probabilities given by the MLP via a product rule is proposed. It can be simply obtained by multiplying the posterior probabilities given by the MLP via a product rule is obtained when the FF features were combined with J-Rasta-PLP features. The J-Rasta technique with the frequency filtering enhanced the robustness of the FF features in noisy conditions. Finally, the power of the multi-stream technique has been used to obtain further improved results with a combination of the FF features with the J-Rasta-PLP features using an improved product rule.

In all the tests, the FF features resulted in a degraded performance for the factory noise. As the spectrum of that noise shows a strongly low-pass characteristic, which is not completely removed by the used band-pass filter, the degradation may be at least partially due to the high noise contamination of the first FF feature, which is an absolute low-pass energy for both FF1 and FF2 filters. On the other hand, the suppression of that first FF feature when the noise has strong energy at low frequencies resulted in a clear improvement of recognition accuracy in a previously reported work [3]. In consequence, additional tests have to be done for both kinds of noise to test the influence of that low-pass energy parameter on the recognition performance.

### Table 7: Combination of df (best FF features with factory noise) and J-Rasta-PLP features, in factory noise conditions

<table>
<thead>
<tr>
<th>WER</th>
<th>SNR = 0</th>
<th>SNR = 6</th>
<th>SNR =12</th>
<th>SNR = 18</th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
<td>63.0</td>
<td>33.9</td>
<td>19.3</td>
<td>12.9</td>
</tr>
<tr>
<td>J-Rasta-PLP</td>
<td>41.8</td>
<td>21.1</td>
<td>12.8</td>
<td>9.7</td>
</tr>
<tr>
<td>Product rule</td>
<td>50.0</td>
<td>24.2</td>
<td>13.4</td>
<td>9.6</td>
</tr>
<tr>
<td>New product rule</td>
<td>43.4</td>
<td>20.6</td>
<td>11.3</td>
<td>8.0</td>
</tr>
</tbody>
</table>

### 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 the first part of the study, we observed that the speech recognition performance of the FF features is better than the performance of the well-known MFCCs in all cases, with a lower computational cost. Moreover, the results of the FF features has been improved by concatenating them with other frequency features.

In the second part of this article, the combination of Rasta technique with the frequency filtering enhanced the robustness of the FF features in noisy conditions. Finally, the power of the multi-stream technique has been used to obtain further improved results with a combination of the FF features with the J-Rasta-PLP features using an improved product rule.

### 8. REFERENCES