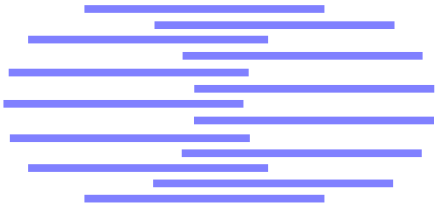
IDIAP RESEARCH REPORT

IDIAP

Martigny - Valais - Suisse



Subband–Based Speech Recognition in Noisy Conditions: The Full Combination Approach

Astrid Hagen † Andrew Morris †
Hervé Bourlard †

IDIAP–RR 98–15

October 1998

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

† IDIAP—Dalle Molle Institute of Perceptual Artificial Intelligence, P.O. Box 592, CH-1920 Martigny, Switzerland, {hagen,morris,bourlard}@idiap.ch.

Subband–Based Speech Recognition in Noisy Conditions: The Full Combination Approach

Astrid Hagen Andrew Morris Hervé Bourlard

October 1998

Abstract.

In this report, we investigate and compare different subband-based Automatic Speech Recognition (ASR) approaches, including an original approach, referred to as the "full combination approach", based on an estimate of the (noise-) weighted sum of posterior probabilities for all possible subband combinations. We show that the proposed estimate is a good approximation of the ideal, but often unpractical, solution consisting in explicitly considering all possible subband subsets. This approximation results in a nonlinear, still simple and easy to implement, combination function. As opposed to other subband-based approaches, we believe that the proposed solution is more optimal (mathematically correct) and allows us to relax some of the (subband) independence assumptions. Similarly to this full posterior combination approach, which combines the subbands after independent processing, a full feature combination approach is investigated, in which all the possible subband features are orthogonalized and combined into a single feature vector (before probability estimation).

The different approaches have been tested and compared on the Numbers'95 database (free format numbers) with different levels of (Noisex'92) car noise. This was done on the basis of two different acoustic features, namely PLP and J-RASTA-PLP features, and different weighting schemes. Those experiments show that the full combination approximation yields very good estimates of the actual full combination posteriors and that both approaches yield very good recognition performance.

Acknowledgements: The support of the OFES under the grant for the "Speech, Hearing and Recognition" (SPHEAR) project # OFES 970299 is gratefully acknowledged. This paper benefited from fruitfull discussions with my colleagues at IDIAP, including Hervé Glottin, Katrin Keller and Christopher Kermorvant, as well as colleagues at other institutes such as Stephan Dupont at FPMs at Mons, Belgium, and Nikky Mirghafori and Brian Kingsbury at ICSI, Berkeley, USA.

Contents

1 Introduction

In this report we investigate different subband-based Automatic Speech Recognition (ASR) approaches in clean and noisy environments. As already introduced previously (see, e.g., [BDR96, DB96, HTP96], the general idea of subband-based ASR is to split the frequency range into several bands, and to use the information in each band for phonetic probability estimation by independent modules. These probabilities are then combined for recognition later in the process at some segmental level. In the present study, we will not attempt to optimize this segmental level but will only focus on combination at the acoustic frame level. In this framework, the present report will discuss and compare different combination paradigms with particular attention to a new scheme based on a noise weighted decomposition of phoneme posteriors. In each case, different subband and fullband acoustic features will also be compared.

Consequently, on top of some theoretical discussions, this report will assess further the potential advantages of subband-based ASR on a difficult task (continuous free format numbers) and real noise (from the Noisex database) conditions. Indeed, the subband approach has many potential advantages, including better robustness to narrow band noise. In this approach, features will be extracted from small subbands and processed independently of each other. Consequently, as opposed to fullband cepstral analysis which typically spreads any noise across all the components of the acoustic vector, it can be expected that a subband approach has greater potential for robustness to noise. However, as a disadvantage, this is often achieved by introducing assumptions regarding the correlation between the subbands and the possible frequency localization of the noise. This leads to the following dilemma. On the one hand, the subband paradigm requires that we work in the spectral domain, in which case subbands are correlated. On the other hand, if we orthogonalize the components of the frequency band (e.g., by using cepstral or other such linear transform), we are no longer in the spectral domain and have already spread the noise if there was any. The limits of this independence assumption, as well as some solutions, will be discussed here.

The work reported here takes place in the framework of hybrid HMM/ANN systems using Artificial Neural Networks (ANN) to generate local emission probabilities (posterior probabilities in our case) for Hidden Markov Models (HMM). Although most of the developments reported here would apply equally to HMM/ANN or standard HMM systems, the hybrid system presents several advantages including e.g., their discriminant properties and the availability of local posteriors (which will be widely exploited in our work). In the first section, the mathematical framework of our solution to the restricting independence assumption, referred to as the "full combination" approach, will be presented. However, since this approach involves the training of a large (usually prohibitive) number of recognition modules (i.e., ANNs in our case), an approximation to this "optimal" approach will be introduced. This approximation will result in an interesting, although simple and easy to implement, nonlinear combination function.[1] For both methods, the combination of the subbands involve weighting factors, the calculation of which will be described in the following section.

Most of the approaches considered here will combine the subbands, after they have been independently processed, at the local probability level. However, it is clear that the subband features could also be combined at the feature level, after feature extraction and orthogonalization, and concatenated to form a single acoustic vector before further processing. This method, referred to as "feature combination", was initially proposed in [OBP98] and also prevents the possible band limited noise from spreading across all acoustic components. This approach was extended to the "full feature combination" approach in which the features of all possible subband subsets are recombined into one single acoustic vector and will be compared to the alternative solutions proposed here.

Finally, various experimental results will be presented on clean and noisy data, i.e. continuous free format numbers (Numbers'95) with additive noise (Noisex'92).

[1] We recall here that subband analysis and *nonlinear* combination is known to take place in the auditory system.

2 Theoretical Framework

After a very brief description of the subband approach previously used, the theoretical framework of the full combination approach and its approximation, as well as the estimation of the weighting factors, will be described.

Let q_k denote a particular class corresponding to one of the ANN outputs (which, in hybrid HMM/ANN systems, will often be associated with a particular phone) and $X = x_1, ..., x_n, ..., x_N$ the acoustic vector sequence associated with a particular utterance, where $x_n = x_{n,1}, x_{n,2}, ..., x_{n,d}$ is a d-dimensional vector or, in the case of subband processing, a vector of d-subband vectors. Since this report will mainly deal with local combination, equations can be written without the temporal index, and we will denote x_n simply as x and $x_{n,d}$ as x_d, and a component or subband subset will be denoted $x_{ij} = \{x_i, x_j\}$. In hybrid HMM/ANN systems, the ANN has been trained to generate local posteriors $P(q_k|x)$, for all possible HMM states q_k $(k = 1, ..., K)$, given x at its input.[2]

2.1 Initial Subband Approach

As illustrated in Figure 1, most of the subband-based ASR approaches developed so far consisted in splitting the frequency range into several bands, pre-processed independently of each other, and in feeding the resulting subband features into independent recognizers. The subband (posterior) phonetic probabilities are then combined for recognition later in the process at some segmental level. Ideally, this approach should consider all possible subband combinations and select the best one, as confirmed by the experiments reported in [HTP96]. However, since first it is not always tractable to consider all possible combinations and second it is very difficult to automatically select the best subband combination, most of the subband approaches use simple combination schemes of a few disjoint subbands, assuming that the frequency bands are independent and that the noise is limited to one of these bands [BDR96, DB96], resulting in the following approximation:

$$P(q_k|x) = \sum_{i=1}^{d} w_i P(q_k|x_i) \tag{1}$$

where, in our case, each $P(q_k|x_i)$ is computed with an ANN with x_i (possibly with temporal context) at its input. w_i are weighting factors. Figure 1 illustrates this for the simple case of two frequency subbands x_1 and x_2.

Usually, in the subband systems used so far, the acoustic vectors are decomposed into a small set of typically 4 [BD97, DB96, DBR97] or 7 [HTP96] subbands. In the work reported here, we only considered the case of 4 subbands, resulting in acoustic vectors x divided into 4 sub-vectors x_i, with $i = 1, ..., 4$.

2.2 Full Combination Approach

In the following, we will now focus on the possible generalizations of (1) and a possible way to actually combine the evidence from all possible subband subsets. Compared to the previous case, and as illustrated in Figure 2 for the case of two subbands x_1 and x_2, we would like to estimate and combine the (posterior) probabilities from all possible subband subsets, i.e., $P(q_k|x_1, x_2) = P(q_k|x_{12})$ (assuming that all bands are reliable), $P(q_k|x_1)$, $P(q_k|x_2)$, and $P(q_k)$ (assuming that none of the subbands is reliable).

In this "optimal" posterior combination scheme, we assume that some unknown components of x are noisy and less reliable. The missing information concerning which subset of subbands of x is the most reliable for recognition is modelled as a latent variable, n. $P(q_k|x)$ is then estimated by integrating over all possible missing values of n, while associating a probability with each possible

[2] If necessary, these local posteriors can also be divided by the prior probabilities $P(q_k)$ as observed on the training data, resulting in local scaled likelihoods $\frac{P(q_k|x)}{P(q_k)} = \frac{P(x|q_k)}{P(x)}$.

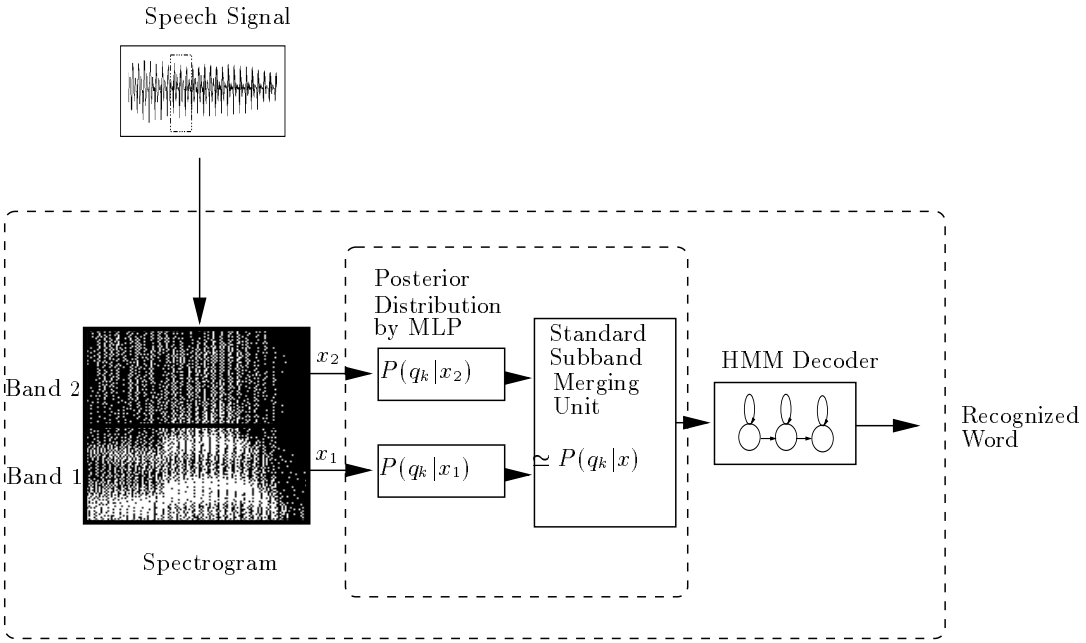


Figure 1: Illustration of the standard multiband-based speech recognition approach on two subbands.

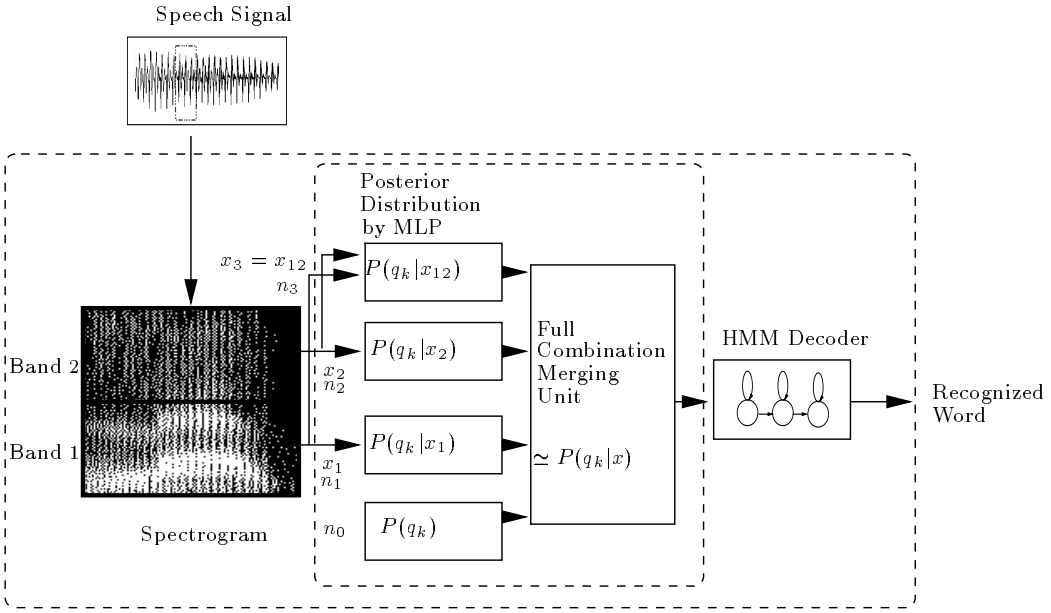


Figure 2: Illustration of the full combination approach on two subbands.

value. If we have acoustic vectors of dimension d (or d subbands), we have $C = 2^d$ possible noisy subsets (including the empty and full sets). Integrating over all possible values of n, and given that the associated possibilities are exhaustive and mutually exclusive, we can write:

$$\begin{aligned} P(q_k|x) &= P(q_k \wedge (n_1 \vee n_2 \vee \cdots n_C)|x) \quad \text{(exhaustive)} \\ &= \sum_{i=1}^{C} P(q_k, n_i|x) \quad \text{(mutually exclusive)} \\ &= \sum_{i=1}^{C} P(q_k|x, n_i) P(n_i|x) \end{aligned} \tag{2}$$

We now consider how to estimate each of the terms in (2). The first term $P(q_k|x, n_i)$ is the probability for state/phoneme q_k given the current acoustic vector x and the knowledge that subset i of x is the most reliable. While this could be estimated in various ways, recent experiments in recognition with missing data [?, MCG98, Her98] have shown that, when the position of highly inaccurate (effectively missing) data is known, recognition can be strongly improved simply by ignoring this data. In the case of recognition based on stochastic models trained on clean data, this can be effected by computing marginal distributions with respect to the reliable data. In our case this corresponds to simply equating the joint information (x, n_i) in (2), which can be read as "data x, of which only subset i of x contains reliable information", with subset i of x, which will be denoted x_{c_i} [3]. Consequently (2) becomes:

$$P(q_k|x) = \sum_{i=1}^{C} P(q_k|x_{c_i}) P(n_i|x) \tag{3}$$

$P(n_i|x_{c_i})$ is then estimated as the probability that all of the data in subset i is clean, and all of the data in its complement is noisy.

We are now left with the following two problems to investigate:

1. How to compute the sum over all the C possible combinations? A direct approach would be, for each term of the sum, i.e. for each possible combination of subbands, to train a separate ANN to estimate each of the probabilities $P(q_k|x_{c_i})$, thus requiring a large, possibly prohibitive, number of neural networks. For example, in the case of 4 subbands, we would have 16 neural networks to train (with different input subsets) and to estimate and combine at every time step to compute $P(q_k|x)$. In Section 2.3, we will propose an approximation that allows us to avoid training and estimating all possibilities, and where every term $P(q_k|x_{c_i})$ required in (3) is estimated on the basis of a minimal set of neural networks, typically the single subband neural networks. For the case of 4 subbands (where it is still possible to train and estimate 16 neural networks), and for different noise conditions, experiments will be shown in which the proposed approximation yields performance acceptably close to the solution of actually combining all possible subband neural nets.

2. How to estimate the weighting factors $P(n_i|x)$ in (3)? These weights, which represent the relative utility for recognition of each subband combination x_{c_i}, are clearly very important for this technique and their estimation will be discussed in section 2.4.

2.3 Approximation of the Full Combination Posteriors

Computation of (3) requires the posteriors $\hat{P}(q_k|x_{c_i})$ for each of K states from the neural networks trained for all $C = 2^d$ different possible combinations of d subbands. In the experiments which follow

[3] c_i will denote the set of indices for subbands in subset i, so that $x_{c_i} = x_i : i \in c_i$

the number of subbands was limited to 4 (resulting in 16 neural networks) so this was feasible, but this approach rapidly becomes impractical as the number of subbands increases.

This problem would clearly be solved if the C subband combination posteriors $\hat{P}(q_k|x_{c_i})$ for each state q_k could be expressed in terms of the single-subband posteriors $\hat{P}(q_k|x_j)$, $j = (1, ..., d)$, alone. While it is not possible to obtain the exact combination posteriors in this way, it is possible to approximate combination posteriors from single subband posteriors, **without assuming full subband independence** (but only conditional independence), by the following procedure (see Appendix A for full derivation):

$$\bar{P}_{ki} = \frac{\prod_{j \in c_i} P(q_k|x_j)}{P^{|c_i|-1}(q_k)} \quad \forall k = 1, ..., K$$

$$\hat{P}(q_k|x_{c_i}) = \frac{\bar{P}_{ki}}{\sum_{l=1}^{K} \bar{P}_{li}} \tag{4}$$

This is the equation we used in the experiments in Section 3. It is often argued that the combination of evidence from separate frequency subbands in the human auditory system is performed according to a nonlinear function. It is interesting to note here that the approximation (4) when substitued in (3) results quite naturally in such a nonlinear function.

2.4 Estimation of the weighting factors

Let's now interpret the factors $P(n_i|x)$ in (3) to see how they can be modelled. $P(n_i|x)$ is the probability that subset i of x is the most usefull for recognition. This is equivalent to saying that it contains the largest selection of clean data. $P(n_i|x)$ is therefore the probability, based on information present in x, that every subband in subband combination i is clean, and every other subband (the components which were disregarded in the computation of $P(q_k|x_{c_i})$ is noisy. On the assumption that the presence of noise in each subband is independent, this probability $P(n_i|x)$ can therefore be approximated as the product of the probabilities for each subband in subset i being clean, and each remaining subband being noisy [4].

$$\hat{P}(n_i|x) = w_{c_i} \simeq \prod_{j \in c_i} P(x_j \ clean) \cdot \prod_{k \notin c_i} P(x_k \ noisy) \tag{5}$$

The factors on the right hand side of (5) can be estimated from a signal-to-noise (SNR) estimator working in each of the d frequency bands. Based on these estimates together with the assumption that noisy data carries no useful speech information we get the probability that a selection c_i describes the best data selection, which we then use as the weighting factor w_{c_i} for the corresponding subbands combination ANN recognizer.

For this, the d SNR estimates $snr[x_j]$ are first calculated, one for each subband. We convert this SNR value $snr[x_j]$ into an estimate of the probability that the band j is clean (a number between 0 and 1, increasing with SNR) by first truncating it to lie within a known range between snr_{min} and snr_{max} according to (6) below, then shifting it so that its minimum value becomes zero and scaling it so that its maximum value becomes 1 according to (7). The SNR range limits are taken as $snr_{min} = 0$ and $snr_{max} = 30$, because below and above these limits it is likely that the data is harmful/useful respectively.

$$snr'[x_j \ clean] = max(min(snr[x_j], snr_{max}), snr_{min}) \tag{6}$$

[4] Note that $\forall$ bands $k : P(x_k \ noisy) = 1 - P(x_k \ clean)$.

$$w_j = P(x_j \; clean) = \frac{snr'[x_j \; clean] - snr_{min}}{snr_{max} - snr_{min}} \tag{7}$$

After that, these probabilities, i.e. the probability of a band j being clean, are multiplied for each band in the respective combination c_i. This is then multiplied by the probability of each band k not in the combination being noisy, resulting in the relative reliability w_{c_i} of combination c_i according to (5).

In the absence of any reliable information about which bands are corrupted by noise, we could assume equal weights for each subband combination. This would correspond to simply forming the average of the output from the multiple recognizers, which is also often found to improve classification performance [Bis95, BD97].

2.5 Feature Combination

Finally, we will also compare the methods described above to an approach, initially proposed in [OBP98] and consisting of recombining subband features after having orthogonalized them locally (e.g. by performing subband cepstral analysis, as initially done in [BDR96, BD96]). In this way, if there is noise in one of the subbands, the cepstral analysis will not spread it in all the acoustic components. In [OBP98], this has been done with a filter bank analysis, followed by a mel-cepstrum analysis[5] applied on each of the subbands individually. The mel-cepstrum vector presented to the classifier is then created by concatenating all the subband mel-cepstrum vectors. Their recognizer was based on AT&T's ATIS[6] speech recognizer [BRA95] using context independent phoneme HMMs with 3 states and 16 mixture Gaussian distributions.

Tests were run for the fullband system and both the conventional likelihood and the new feature combination system on noisy speech. Different noises from the Noisex'92 database were added to the clean speech of the ARPA ATIS continuous speech database. The feature combination system gave better performance than the likelihood combination system for all noise conditions (babble, factory, machine gun, car noise, etc.). Both systems performed almost always better than the fullband system.

In the case of car noise, which was added at 10 dB SNR, their fullband system resulted in a 9.0% word error rate, and the best likelihood combination system in 10.0%, while the feature combination system of 2 subbands resulted in only 8.8% error rate.

Here, and following the spirit of the full combination above, we will do the same thing, but using all the C possible subband combinations. In this case, for each of the C subband sets, we perform cepstral analysis (see Section 3.7 for experimental details and parameter setting) and recombine the resulting C feature vectors into a large feature vector which will be presented at the input of a neural network for training and local probability estimation.

For our subband system of 4 subbands, we have C=14 possible subband combinations plus the fullband domain for feature extraction. As done with the full posterior combination approach, we could also add the priors (posterior estimates in the case when all the data is noisy) to the output of the full feature combination neural network. This was, though, not done in this report. The feature components used in the full feature combination experiments correspond exactly to the set of PLP- and respectively J-Rasta-PLP features as they were extracted for the training and testing of the full posterior combination ANNs in Section 3.2.

3 Experiments

In the following sections, the experimental setups, including the databases for the speech and noise data and the feature extraction, as well as the different trained full- and subband recognizers will be described. Then, the results of the various experiments incorporating these recognizers will be discussed.

[5] Based on the **D**iscrete **C**osine **T**ransform.
[6] **A**ir **T**ravel and **I**nformation **S**ervice.

3.1 Numbers'95 and Noisex'92 Databases

The Numbers'95 database [CNLD95] is a release of the Center for Spoken Language Understanding (CSLU), which collects databases over the telephone making them available to academic institutions. The databases are well labeled by experienced transcribers following the CSLU Labeling Conventions [LM94]. This way it is possible to objectively compare performance results on these databases among competing institutes. For this reason, we chose the Numbers'95 database which consists of many kinds of connected digit sequences: cardinal numbers, ordinal numbers, and digit strings, like for example people saying their house number or zipcode (US) when giving their address. Numbers which appeared in the middle of a sentence were extracted from the host utterance and saved in separate files to make up the Numbers'95 corpus. As a whole, the database consists of 10.5 hours of speech.

For the experiments, the database is split into three major parts: the training set, including the cross validation (CV) set, and the test set. The training set consists of 3590 phrases, of which 10% (357 utterances) are used for cross-validation. The remaining set of 3233 utterances is used to train the respective full- and subband-ANNs. After each cycle through the training set, it is examined on the cross-validation set to what extent the ANN is already trained. If the error rate of the respective ANN did not further diminish after the previous evaluation on the CV set, the training is stopped. The test set comprises 1227 utterances and is used to test newly developed recognizers. For the preliminary results and to speed up the experiments, only the first 100 test sentences were used here. Moreover, the whole frequency domain was used for feature extraction and processing, i.e. we did not only concentrate on the often chosen frequency domain of telephone speech from environ 300 Hz to 3300 Hz. This way, more noise might have been introduced in the speech-free frequencies.

For the experiments with corrupted speech, we chose the Noisex'92 database [VSTJ92]. The noise section of the Noisex'92 database consists of eight different noises, such as e.g. car noise, factory noise or machine gun noise, providing disparate difficulties for recognizers as for example non-stationary versus stationary noise. Noise was sampled at 16 kHz and, thus had to be down-sampled before adding to our Numbers'95 data which is sampled at 8 kHz.

In the framework of our SPHEAR project[7], project partners of which are the Universities of Bochum, Grenoble, Keele, Patras and Sheffield as well as Daimler Benz, we chose car noise[8] to corrupt the clean data of the Numbers'95 test set. Different signal-to-noise ratios (SNR) were employed, ranging from -30 to 30 dB. The SNRs were estimated using the STRUT[9] programme "estimate-snr-cbe", which is based on the energy histogram clustering method [BD98]. The respective SNRs are calculated in the selected frequency (sub-)bands. After the calculation of the SNR for each file frame by frame, the noise was added to the test set at the required global SNR. The training set was left clean.

3.2 General Setup

3.2.1 Feature Extraction

The first set of experiments was carried out on PLP[10] features [Her90], to evaluate the new approach on a first set of well-known and well-performing features. (PLP analysis was described by Hermansky [Her98] as the most efficient speech representation according to an extensive DARPA evaluation).

The second set of experiments incorporated a set of features that are proven to be more noise robust: the J-Rasta-PLP features [HM94].

In the PLP-technique [Her90], properties of human hearing are simulated to gain an auditory spectrum by convolving the FFT spectrum with the critical-band function, multiplying by a fixed

[7] See www.dcs.shef.ac.uk/~pdg/sphear/sphear.htm and www.idiap.ch/~kermorva/Sphear/sphear_index.html
[8] Car noise number 23.ns of the Noisex'92 database.
[9] For the STRUT "homepage" please see: http://tcts.fpms.ac.be.
[10] PLP stands for **P**erceptual **L**inear **P**rediction.

equal loudness curve, and compressing its amplitude by a cubic-root function. The auditory spectrum is then approximated by an autoregressive all-pole model to smooth out irrelevant details and introduce temporal context.

To render the features more noise robust, other features than the spectral PLP-features were considered. For the Rasta[11]-PLP-features, the speech signal is, moreover, filtered before being processed as described above for the PLP-features. Each frequency channel is band-pass filtered in such a way that any constant or slowly varying components in the frequency channel are removed, resulting in spectral estimates which are less sensitive to slow variations in the short-term spectrum as can be introduced by a communication channel. Thus, these (log-)Rasta-PLP features are mainly robust to convolutional noise, as e.g. distortions introduced by the telephone line.

On the other hand, uncorrelated additive noise components cannot be effectively removed by the Rasta band-pass filtering in the logarithmic domain. For this, the J-Rasta-PLP features were introduced [HM94]. Depending on the noise level, the variable J is altered to introduce the Rasta processing scheme in different spectral domains for each noise level: for small spectral values a linear-like domain is chosen, for large spectral values a logarithmic like domain. This renders the J-Rasta-PLP features more robust to both additive and convolutional noise than the (log-)Rasta-PLP features and much more robust than the simple PLP features themselves.

3.2.2 Subband- and Fullband-MLP Structures

Our experiments were carried out in the framework of hybrid HMM/ANN-systems using artificial neural networks to generate the local posteriors for the HMMs. The ANNs we used are Multilayer Perceptrons (MLPs) with one hidden layer of 1000 hidden units. The output units corresponded to the 33 phonemes (one output unit for each phoneme) representing the lexicon, and were processed by an HMM decoder based on single state phonetic models with duration modeling. The features, which will be described in the following, were extracted every 12.5 ms on a window of 25 ms. The topologies of both the fullband and the subband systems stayed the same for all experiments, only varying in the number of input units according to best results as found in earlier experiments.

As described in Section 2.2, one MLP was trained for each possible subband combination, resulting in 15 MLPs. For the case of the MLP which would have to be trained on no data, we used the priors which were extracted from the training set and which constitute in this case the posterior probabilities. Thus, we have 16 posteriors for each phoneme for each frame, which were then recombined in two different ways.

In a first test (Test I), the posterior probabilities of the 16 MLPs were combined by a weighted sum with equal weights, which corresponds to the arithmetic mean. In Test II, the MLP outputs were combined using a weighted sum of SNR-based weights as introduced in Section ??. As a short reminder for the reader we would like to recall here that the SNR-values were calculated automatically and not taken a priori.

MLPs on PLP-Features

The fullband HMM/MLP-hybrid system for our first experiments using PLP-features [Her90] consisted of an MLP with 351 input units (9 frames of 39 features). The input layer of each (sub-band and fullband) MLP had a context window of nine frames – one current frame, four frames into the past and four frames into the future. The features were 13 fullband PLP-features including energy, as well as their first and second order derivatives (summing up to 39 features).

The features used for each subband recognizer were 9 subband PLP-features (compare Table 9 in Appendix B), including energy, complemented by their first and second order temporal derivatives. With 9 frames of contextual information at the input of the MLPs this resulted in 243 input units.

[11] RASTA is the acronym for **R**el**A**tive **S**pec**T**r**A**l methodology [HMBK92].

MLPs on J-Rasta-PLP Features

Since it is known that the standard multi-band approach [BDR96] is less efficient than other noise cancellation techniques in the case of narrowband [BD96] and wideband noise [HM94, Hir93], we used J-Rasta-PLP features for the second set of experiments. The subband-PLP features were therefore independently computed for each subband (combination), previously processed with the J-Rasta filter (for detailed description of the feature extraction parameters, please see Table 10 in Appendix C). The J-value was adapted automatically according to noise-power. Again, the features' first and second order derivatives including the energy were added to the feature vector. This resulted in MLPs of 243 input units (9 frames of 27 features).

The MLP trained on the full frequency domain consisted of 351 input units, comprising 9 times 13 J-Rasta-PLP features, including energy, as well as their first and second order derivatives.

MLPs on Recombined Features

For the experiments on the recombined feature vector, the subband PLP- and J-Rasta-PLP features described in the paragraphs above were recombined respectively into one single feature vector of 144 elements. These were taken over 5 frames of contextual input for the training of the ANNs.
The feature combination ANNs were trained on the PLP- and the J-Rasta-PLP features without (delta-)delta features. As the size of the ANNs were already quite high not more contextual input than the 5 frames and no (delta-)delta features could be used.
Only in the case of the J-Rasta-PLP features did we train another ANN also including the delta-features as the results of the first ANN (without delta-features) already seemed very promising. This training almost took one week and it therefore does not seem feasible to also include the delta-delta features although this might still improve the recognition performance. Thus, for the ANN on the PLP-features we had 720 input units, and for the ANNs on the J-Rasta-PLP features 720 and 1440 input units respectively. The hidden layer comprised 1500 units for the ANNs of 720 input units and 3000 for the ANN of 1440 input units. The output layers consisted of the 33 units for the phoneme classes.

3.3 Fullband Experiments

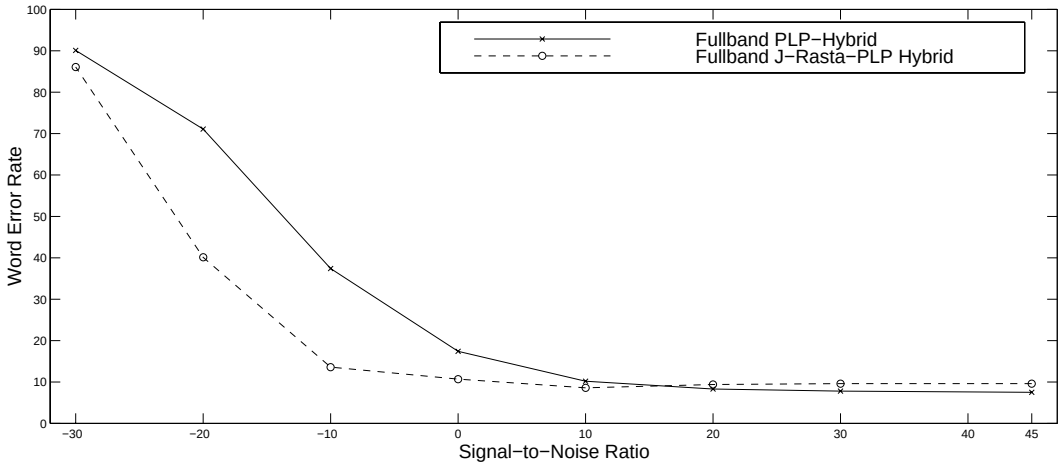


Figure 3: Illustration of the Results of the two Fullband HMM/ANN-Hybrid Systems.

First, experiments were carried out on the two fullband MLPs, one for the PLP-features and one

for the J-Rasta-PLP features. Tests were run on clean and noise-added speech from -30 to +30 dB SNR. The results can be seen in Table 1 and Figure 3. For clean speech the fullband MLP trained on fullband PLP thirteenth-order features (plus deltas and delta-deltas) gave best results (7.5 % Word Error Rate (WER)). The fullband MLP trained on fullband J-Rasta-PLP nineth-order features (plus deltas and delta-deltas) resulted in a slightly worse WER of 9.6 %.
Looking at the noisy data it can be seen that the system with the PLP-features (middle row of Table 1) is less noise robust and already degrades at 0 dB SNR. The MLP trained on the J-Rasta-PLP features (last row in Table 1) stays robust down to -10 dB SNR and starts only then to deteriorate in performance (at -20, -30 dB SNR).

System	Signal-To-Noise Ratio							clean
	-30	-20	-10	0	10	20	30	45
PLP-features	90.1	71.1	37.4	17.4	10.2	8.3	7.8	7.5
J-Rasta-PLP features	86.1	40.1	13.6	10.7	8.6	9.4	9.6	9.6

Table 1: Experiments on the Fullband HMM/MLP-Hybrid Systems.

3.4 Initial Subband Experiments

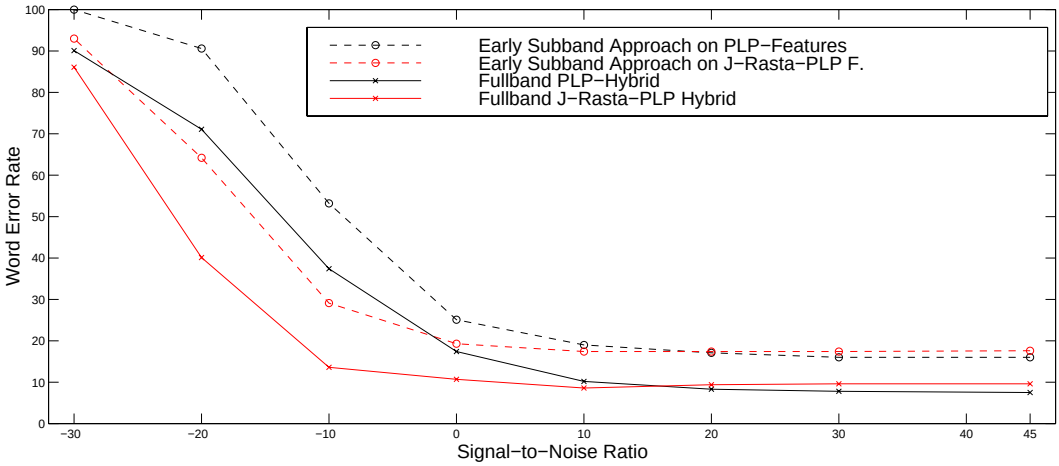


Figure 4: Illustration of the Results of the two Fullband HMM/ANN-Hybrid Systems compared to the Results of the Standard Subband Approach.

As pointed out earlier, in the standard subband approach the subband recognizers were each trained on a single subband only, as illustrated in Figure 1. The output probabilities of all recognizers were then recombined in a (possibly weighted) sum without further trying to approximate missing combinations of subbands. In order to see the improvement we might gain with the new, full combination approach as compared to the standard subband approach, we present in this section the results of the standard subband approach. The tests were run with the same MLPs as were described above, but only using the 4 MLPs for subbands 1, 2, 3 and 4. Their output (i.e. the posterior probabilities from each subband MLP) was combined by a sum of equal weights. Results can be seen in Table 2 and Figure 4. In Figure 4 they are compared to the results we gained on the respective fullband systems as described in Figure 3.
Again, the results on the PLP-features without J-Rasta filtering show that these features are less noise robust then the J-Rasta-PLP features, although both systems achieved similar results in the case of

System	Signal-To-Noise Ratio							clean
	-30	-20	-10	0	10	20	30	45
PLP-features	100.0	90.6	53.2	25.1	19.0	17.1	16.0	16.0
J-Rasta-PLP features	93.0	64.2	29.1	19.3	17.4	17.4	17.4	17.6

Table 2: Initial Subband-Experiments with Standard Combination of 4 MLPs only (no approximation of the missing combinations).

clean speech. As compared to the fullband systems, the results of the initial subbands approach on neither the PLP-features nor the J-Rasta-PLP features could approximate those of the fullband systems. This result was observed for both clean and noise-added speech.

3.5 Full Combination

In the following sections, we present the experiments and results for the full combination approach and its approximation.

Full Combination on PLP-Features

Resulting word error rates for the full combination approach on the PLP-features for clean and noisy data are reported in Table 3. Recognition performance of the pure fullband system for different SNRs are compared with the differently weighted (equal weights and SNR-based weights) method for combining all possible subband combinations.

System	Signal-To-Noise Ratio							clean
	-30	-20	-10	0	10	20	30	45
Fullband Hybrid	90.1	71.1	37.4	17.4	10.2	8.3	7.8	7.5
16 MLPs, equ. weights	93.0	68.2	34.5	15.0	11.5	9.9	9.1	9.4
16 MLPs, SNR weights	96.3	67.1	31.3	15.5	11.2	8.8	8.8	9.1

Table 3: Subband-PLP features and Fullband-PLP: full combination of the 16 MLPs (Test I with equal weights, Test II with SNR weights) versus the fullband hybrid.

As can be seen in Table 3 and Figure 5, for the PLP-features the word error rate could for low SNR values be improved by the new full combination approach, already when using equal weights only. For almost clean speech and speech with positive SNR (10–30 dB) the fullband system seems to be slightly better. But starting at 0 dB SNR down to -20 dB SNR the combination of the 16 MLPs achieved clearly better results than the fullband system. For the worst case of noisy speech (-30 dB SNR), all systems had almost the same high error rate.

The last row in Table 3 shows the results for the sum with SNR-conditioned weights as described in Section ??. The SNR weights, carefully computed for every subband combination as described in Section ??, further improved the combinations system, although only a very simple noise-estimation technique was used. The error rates for the full combination systems decreased for the (almost) clean data (45 and 30 dB SNR) as compared to the equal weights case. For SNR of 20 dB the word error rate could even be improved by almost one percent point. Looking at SNR values from 0 to -20 dB, the error rate was clearly decreased compared to the fullband system with a little improvement as compared to the equal-weights case in the second row for SNR of -10 and -20 dB. Only for -30 dB SNR no improvement could be achieved.

We are planning on investigating other, more powerful noise estimation (and weighting) strategies (cf. Section 4) to further improve the full combination systems with SNR-weights as compared to the fullband system.

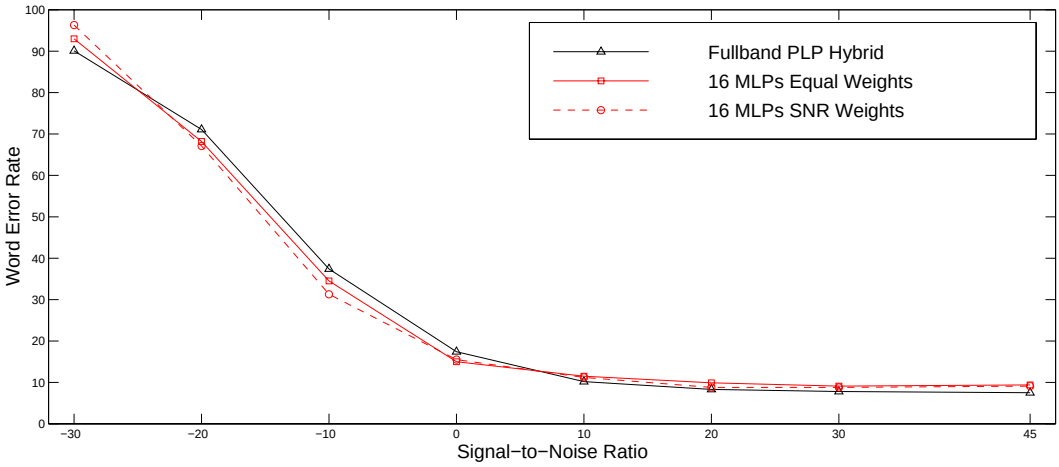


Figure 5: Subband-PLP features and Fullband-PLP: Illustration of all 16 possible subbands combination (equal weights and SNR weights) versus the fullband hybrid.

Full Combination on J-Rasta-PLP Features

In the case of the J-Rasta-PLP features the experiments gave the following results (cf. Table 4 and Figure 6). The full combination subbands approach with the sum of equal weights seems to offer no advantage, but it does not deteriorate the system either (Test I). In the case of the SNR-weighted sum of combinations (Test II), the error rate was clearly improved for SNR values from 45 dB to 20 dB, even compared to the fullband system. The best yielded word error rate on clean speech in this article for the J-Rasta-PLP features was achieved by this system (8.3%, as compared to 9.6% of the fullband hybrid). For more noisy speech (SNR 0 to -30 dB) no improvement could be achieved, but even deterioration at -20 dB SNR.[12]

System	Signal-To-Noise Ratio							clean
	-30	-20	-10	0	10	20	30	45
Fullband Hybrid	86.1	40.1	13.6	10.7	8.6	9.4	9.6	9.6
16 MLPs, equal weights	91.7	44.9	14.4	10.7	10.7	11.0	11.0	11.0
16 MLPs, SNR weights	90.6	56.4	15.5	11.0	9.1	8.8	8.6	8.3

Table 4: Subband J-Rasta-PLP features and Fullband J-Rasta-PLP: full combination of 16 MLPs versus the fullband HMM/MLP-Hybrid System.

The J-Rasta-PLP features are by themselves already more noise robust than the PLP-features as can be seen in the performance of the two fullband systems (first rows in Tables 3 and 4). But this is also true for the subband combination systems. The incorporation of the J-Rasta preprocessing filter before feature extraction does not harm the multiband system but renders it as robust as the fullband system. Both systems stay relatively robust even for car noise conditions of down to -10 dB SNR. Only with -20 to -30 dB SNR do both systems significantly degrade in performance.

These results motivate further experiments with the full combination multiband system on other, more challenging noise types, which will be more difficult to cancel out even for the J-Rasta technique. This will further investigate the robustness of the new approach compared to a standard fullband system.

[12] The word-entrance-penalty (wep) for the experiments was not adapted but chosen to approximate best results for all systems, which was found at wep=5.

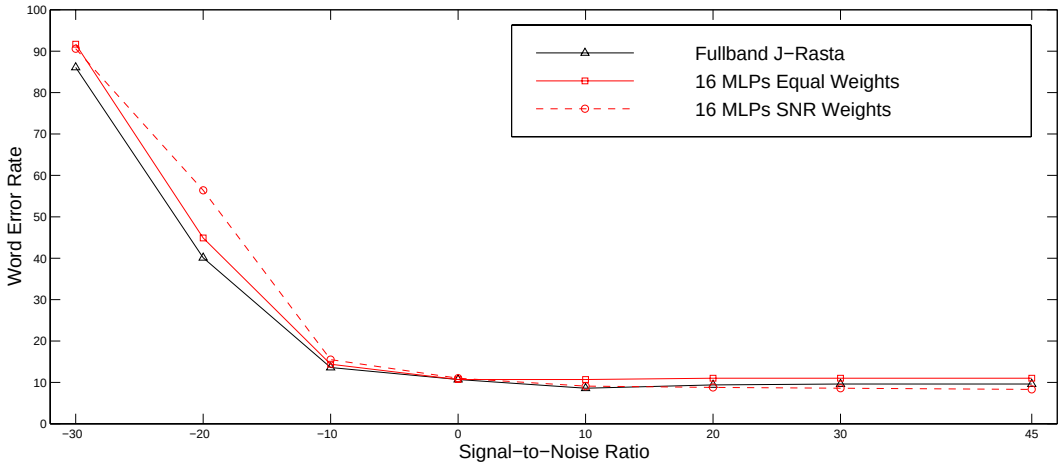


Figure 6: Subband J-Rasta-PLP features and Fullband J-Rasta-PLP features: Illustration of the results for the fullband system as well as the weighted sums of the 16 MLPs (equal weights and SNR weights).

Although these results are convincing and promising for future work in this direction, the reader might be sceptical having to train a large number of MLPs. Therefore, in the following section, we are going to show that with an *appropriate* approximation it is however possible to model this full combination method by a common multi-band HMM/MLP-hybrid system with 4 subbands and 4 trained MLPs only.

3.6 Approximation of the Full Combination (FC) Posteriors

In this experiment we used the already trained subband MLPs from Section 3.2.2 but restricted to the subband recognizers for the isolated subbands 1, 2, 3 and 4 only; the posteriors for all the other subband combination MLPs were obtained using the approximation described in Section 2.3.

The 11 approximated and the 4 original posterior probabilities, as well as the class posteriors of the fullband MLP, resulting in 16 posteriors for each class, were then recombined for the final decision task by a weighted sum of equal weights (Test I) and by a weighted sum of SNR-conditioned weights (Test II).

Approximation of the FC Posteriors on the PLP-Features

System	Signal-To-Noise Ratio							clean
	-30	-20	-10	0	10	20	30	45
Fullband Hybrid	**90.1**	71.1	37.4	17.4	**10.2**	**8.3**	**7.8**	**7.5**
16 MLPs, equ. weights	93.0	68.2	34.5	**15.0**	11.5	9.9	9.1	9.4
16 MLPs, SNR weights	96.3	67.1	**31.3**	15.5	11.2	8.8	8.8	9.1
4 MLPs, equ. weights	92.8	**66.0**	33.7	16.3	10.7	10.2	9.9	9.9
4 MLPs, SNR weights	95.5	**66.0**	33.2	15.8	10.4	9.4	9.9	9.9

Table 5: Subband-PLP and Fullband-PLP features: Approximation by 4 trained MLPs versus the fullband and the full combination approach of 16 MLPs.

Looking at the approximation of the full combination by the 4 MLPs trained on the PLP-features (Table 5 fifth and last row as well as Figure 7), the word error rate on speech of SNRs from 20 to

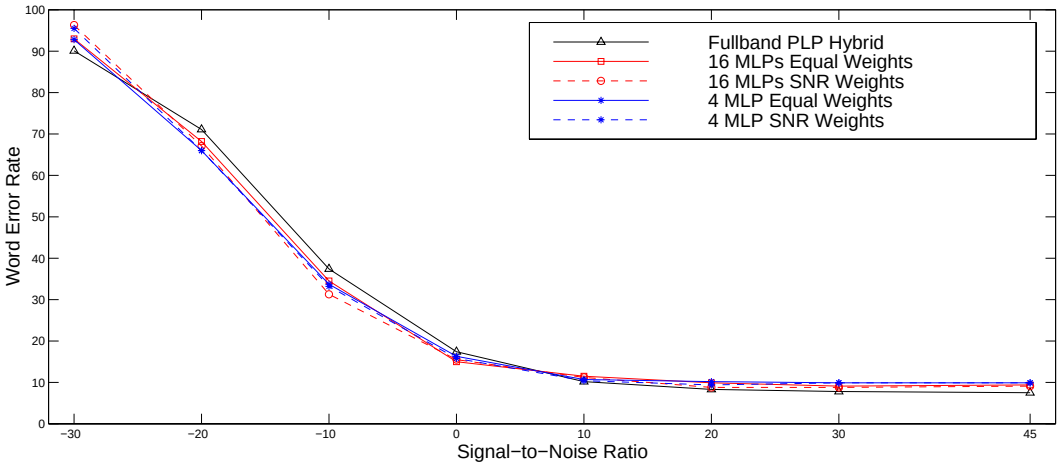


Figure 7: Subband-PLP and Fullband-PLP features: Illustration of the approximation by 4 trained MLPs versus the fullband hybrid and the full combination approach of 16 MLPs.

45 dB achieved about the same results as the full combination approach with all 16 MLPs and, thus a little worse than the fullband system (Test I). In a second test (Test II), the SNR-weighted sum was used, and the results from the approximation by the 4 MLPs were further ameliorated for this range of SNRs, coming closer to the results of the 16 MLPs and the fullband system. In the case of 10 dB SNR the approximation of the full cobmination even gained better results than the real full combination and now shows only little difference as compared to the fullband system.

For SNR rates of 0 dB and -10 dB, it can be seen that the error rate of the 4 MLP system is significantly lower than that of the fullband system. For Test II the improvement could even be increased, rendering the 4 MLPs system almost as good as the 16 MLPs and clearly better than the fullband system.

For -20 dB SNR the error rate for the sum of equal weights (Test I) of the 4 MLPs deteriorated significantly less than for the fullband and the 16 MLPs systems. It could not further be improved with the SNR-weighting (Test II). For SNR of -30 dB all systems degraded in almost the same way. At this SNR level, the SNR-estimation obviously got very difficult so that no improvement with SNR-based weighting could be achieved for neither the 16 MLPs nor the 4 MLPs system.

Approximation of the FC Posteriors on the J-Rasta-PLP features

We now come to the 4 single subband-MLPs trained on the J-Rasta-PLP features. Again, the same MLPs for the four subbands were used in the experiments as had been trained for the experiments of the full combination approach (cf. Section 3.2 and Section 3.5).

For Test I (equal weights) the 4 MLPs approach yield almost the same results as the 16 MLPs approach for SNRs of 10 to 45 dB (cf. Table 6 and Figure 8) and, thus a little worse than the fullband hybrid system. When the SNR-weighted sum is used (Test II), the word error rate can be improved, but not as much as for the full combination approach. Thus, the 4 MLPs did not manage to beat the results of the fullband system or the full combination system of the 16 MLPs.

For SNR values of 0 to -30 dB, the 4 MLPs achieved slightly higher error rates than the full combination approach and with that also higher error rates than the fullband system. Incorporating SNR weights in the sum instead of taking equal weights (Test II), the word error rate could be decreased in some cases (0, -30 dB SNR) but not for all (-10 and -20 dB SNR). The fact that the error rate increased when using the SNR-weighted sum was also observed for the 16 MLPs on SNRs of -20 and -10 dB but only in the case of J-Rasta-PLP features. This has to be investigated. For the

PLP-features both systems had almost always smaller or equal error rates using the SNR-weighted sum as compared to the sum of equal weights.

System	Signal-To-Noise Ratio							clean
	-30	-20	-10	0	10	20	30	45
Fullband Hybrid	**86.1**	**40.1**	**13.6**	**10.7**	**8.6**	9.4	9.6	9.6
16 MLPs, equ. weights	91.7	44.9	14.4	10.7	10.7	11.0	11.0	11.0
16 MLPs, SNR weights	90.6	56.4	15.5	11.0	9.1	**8.8**	**8.6**	**8.3**
4 MLPs, equ. weights	90.1	49.2	18.2	12.3	11.0	11.2	11.2	11.2
4 MLPs, SNR weights	90.6	58.8	21.1	11.0	10.4	10.4	10.4	10.4

Table 6: Subband J-Rasta-PLP and Fullband J-Rasta-PLP features: 4 trained MLPs versus the fullband and the full combination approach of 16 MLPs.

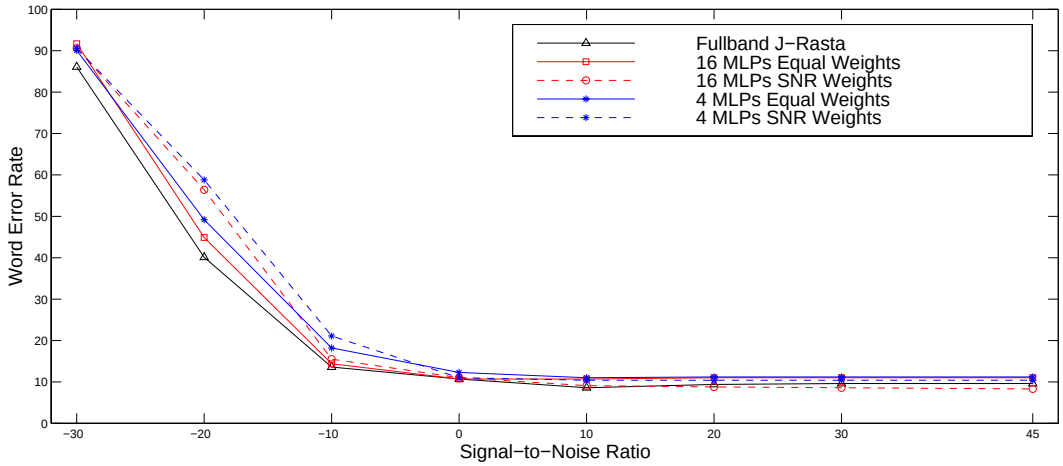


Figure 8: Subband J-Rasta-PLP and Fullband J-Rasta-PLP features: Illustration of the approximation by 4 trained MLPs versus the fullband hybrid and the full combination approach of 16 MLPs.

To sum up, the full combination approach resulted in slightly better performance on clean speech than the fullband system when the J-Rasta features were used but could not further improve performance on the other SNR conditions. The combination approaches using the PLP-features, on the other hand, improved the fullband system on noisy data (0 to -20 dB SNR) even when using the approximation by the 4 MLPs approach only.

3.7 Full Feature Combination

In this section we are going to present the experiments for the full feature combination approach as introduced in Section 2.5.

In [OBP98], it was shown that the feature combination approach yielded better results than their posterior combination strategy. It has to be emphasized again that this was a subband system corresponding to the initial subbands approach as described in Section 2.1. Thus, for a system of 4 subbands, only 4 feature vectors had been extracted and recombined. The results were accordingly compared to their (standard) posterior combination system of 4 MLPs. In our case corresponding to the full posterior combination we also used full feature combination. However, as opposed to [OBP98], we could not achieve better results with the full feature combination approach than with the corresponding

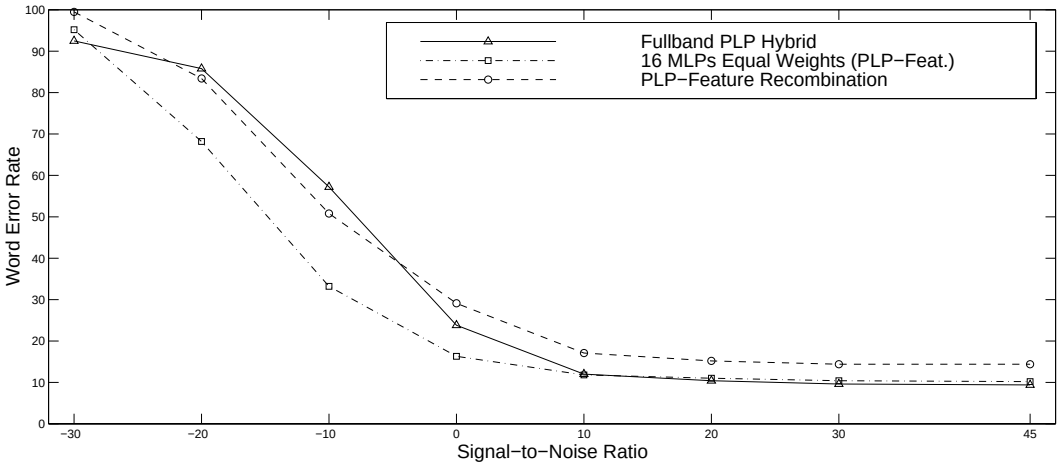


Figure 9: Illustration of the *feature* (PLP-features) combination approach compared to the corresponding fullband hybrid system and the 16 MLPs' *posterior* combination approach.

System	Signal-To-Noise Ratio							clean
	-30	-20	-10	0	10	20	30	45
Fullband Hybrid	92.5	85.8	57.2	23.8	12.0	10.4	9.6	9.4
16 MLPs, equal weights	95.2	68.2	33.2	16.3	11.8	11.0	10.4	10.2
Feature Combination	**99.5**	**83.4**	**50.8**	**29.1**	**17.1**	**15.2**	**14.4**	**14.4**

Table 7: Subband-PLP and Fullband-PLP features: The 15 subband feature combination approach on subband-PLP features compared to the fullband hybrid system and the 16 posteriors combination approach (equal weights).

full posterior combination approach (with the exception of the case of -20 and -30 dB SNR on the J-Rasta-PLP features, see below).

One reason for this could lie in the fact that no delta-delta features (and for the PLP-features no delta-features either) as well as not enough hidden units could be used, as it would not have been possible to train such a big MLP. For the same reason, not the same amount of contextual input could be incorporated as was used for the full posterior approach. Therefore, it is unfortunately rather difficult to compare the results of the full posterior combination approach with those of the full feature combination approach.

The results of the experiments are reported in Table 7 for the PLP-features and in Table 8 for the J-Rasta-PLP features. Still, the full feature combination approach can be compared to the standard fullband hybrid system also trained on the same features (cf. Section 3.3) (first rows in Tables 7 and 8).

For the PLP-features (Table 7), it can be seen that for some noisy data (-10 and -20 dB SNR) the feature combination system (last row in Table 7) achieved better results than the fullband hybrid system, as also illustrated in Figure 9. But as compared to the full posterior combination system, no improvement was gained.

Looking at the J-Rasta-PLP features in Table 8 and Figure 10, we can see that including the delta-features (last row of Table 8) in training and testing improved the recognition rate for clean and all noise conditions as compared to the feature combination system without delta-features (last but one row). For very noisy speech of -20 and -30 dB SNR the full feature combination system with delta-features did even result in smaller error rates than the full posterior combination system. For the other

System	Signal-To-Noise Ratio							clean
	-30	-20	-10	0	10	20	30	45
Fullband Hybrid	86.1	40.1	13.6	10.7	8.6	9.4	9.6	9.6
16 MLPs, equal weights	91.7	44.9	14.4	10.7	10.7	11.0	11.0	11.0
Feature Combination	**88.8**	**51.9**	**26.2**	**17.6**	**15.2**	**13.9**	**13.9**	**13.9**
Feature Combination (incl. Δ's)	**89.0**	**42.5**	**20.3**	**14.4**	**12.6**	**12.0**	**12.0**	**12.0**

Table 8: Subband J-Rasta-PLP and Fullband J-Rasta-PLP features:The 15 subband features combination approach on subband J-Rasta-PLP features (with and without delta-features) compared to the fullband hybrid system and the 16 posteriors combination approach (equal weights).

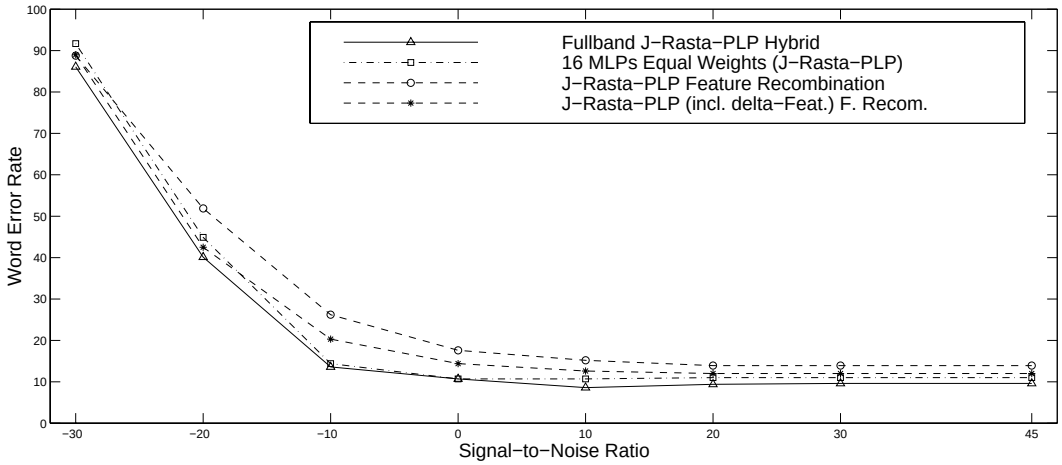


Figure 10: Illustration of the *feature* combination approach (J-Rasta-PLP features) compared to the corresponding fullband hybrid system and the 16 MLPs' *posterior* combination approach.

noise conditions and the clean data, the posterior combination approach (equal weights)[13] was better than the feature combination approach. Compared to the fullband system both feature combination systems (with and without the delta-features) were not able to achieve better recognition results.

4 Conclusion

In this report, we have proposed a new subband based ASR approach allowing us to approximate the full combination of all possible subband subsets, and we compared it with the "optimal" (but often unpractical) full combination approach, as well as other subband ASR systems. The approximation is based on the fact that the fullband posterior probabilities can be expressed as a weighted sum of posterior probabilities, in which each term can also be approximated by a scaled product of available subband posteriors. The weighting can be carried out either with equal weights or with SNRs converted to appropriate weights for each subband combination, the latter usually yiedling better and promizing results.

The experiments were carried out on two sets of features: PLP-features and J-Rasta-PLP features. For both sets of features, and based on 4 subbands, the exhaustive set of 16 MLPs (including the

[13] We chose to compare the full feature combination to the full posterior combination for the case of equal weights only, as this seems more appropriate than considering the SNR-weighted posterior combination. For this reason, the fullband system in Table 8 seems superior to the full posterior combination which is not the case for all SNR-values when SNR-weights were used.

"untrained MLP", only generating class prior probabilities) has been trained to provide a reference point to our approximation. The full combination approach, and the proposed approximation (based on 4 MLPs only) were tested and evaluated on the Numbers'95 database with different levels of additive car noise from the Noisex'92 corpus. In all cases, experimental results showed that, compared to the "optimal" combination, our approximation was always yielding comparable performance. Furthermore, in the case of PLP features, both approaches showed – in the case of noise-added speech – a clear improvement over the standard fullband ASR systems, as well as the "initial" subband approaches tested. Using J-RASTA features though, the conclusions were more mixed. On high SNR values, the recognition performance was slightly (though perhaps not significantly) better for the full combination system, consisting of the 16 MLPs. For small SNR values, all three systems deteriorated approximately the same way, with the fullband system staying slightly more robust than the full combination system and its approximation.

A Derivation of Equation (4)

Let b_i be the number of subbands in x_{c_i} and denote the b_i subbands of x_{c_i} as $y_1, ..., y_{b_i}$. We then have:

$$\begin{aligned} P(q_k|y) &= \frac{P(y|q_k)P(q_k)}{P(y)} \\ &= \frac{P(q_k)}{P(y)} \prod_{j=1}^{b_i} P(y_j|y_1, ..., y_{j-1}, q_k) \end{aligned} \tag{8}$$

We now assume that the subbands are independent when conditioned on a particular class q_k: Spectral continuity guarantees that neighboring spectral bands j and $j-1$ are highly dependent. However, the data in each band will also be close to the mean for any given phoneme q_k. This means that the information which (x_{j-1}, q_k) carries about x_j is not much more than the information carried by q_k alone. The assumption: $P(y_j|y_1, ..., y_{j-1}, q_k) \simeq P(y_j|q_k)$ is far less inaccurate than the assumption: $P(y_j|y_{j-1}) = P(y_j)$. This leads to the following approximation:

$$\begin{aligned} \hat{P}(q_k|y) &\simeq \frac{P(q_k)}{P(y)} \prod_{j=1}^{b_i} P(y_j|q_k) \\ &= \frac{P(q_k)}{P(y)} \prod_{j=1}^{b_i} \frac{P(q_k|y_j)P(y_j)}{P(q_k)} \\ &= \frac{P(q_k)}{P^{b_i}(q_k)} \frac{\prod_{j=1}^{b_i} P(y_j)}{P(y)} \prod_{j=1}^{b_i} P(q_k|y_j) \\ &= \Theta P^{1-b_i}(q_k) \prod_{j=1}^{b_i} P(q_k|y_j) \\ &= \Theta P^{1-b_i}(q_k) \prod_{j \in c_i} P(q_k|x_j) \end{aligned} \tag{9}$$

Because Θ is constant $\forall q_k, k = 1, ..., K$, it will be elliminated by normalizing posteriors $P(q_k|x_{c_i})$ to sum to one.

$$\begin{aligned} \hat{P}(q_k|x_{c_i})/\Theta &= P^{1-b_i}(q_k) \prod_{j \in c_i} P(q_k|x_j) \\ \hat{P}(q_k|x_{c_i}) &= \frac{\hat{P}(q_k|x_{c_i})}{\sum_{l=1}^{N} \hat{P}(q_l|x_{c_i})} \frac{\Theta}{\Theta} \\ &= \frac{P^{1-b_i}(q_k) \prod_{j \in c_i} P(q_k|x_j)}{\sum_{l=1}^{N} P^{1-b_i}(q_l) \prod_{j \in c_i} P(q_l|x_j)} \end{aligned} \tag{10}$$

B Description of PLP–Features for the Different Subband Combinations

In the following Table 9 the parameters for the calculation of the PLP–Subband Features are summarized, which were used by the STRUT programme 'lpc-cepstrum-bands' and 'lpc-cepstrum-several-bands'[14]. The features were Rasta–filtered before processing into the different subbands.

bands	lpc–cepstrum–bands									
	coeff.	lpc order	1^{st} cb[a]	(Hz)	last cb	(Hz)	1^{st} cb2[b]	(Hz)	last cb2	(Hz)
1	8	5	2	(100)	8	(920)				
2	8	3	8	(770)	12	(1720)				
3	8	3	12	(1480)	15	(2700)				
4	8	3	14	(2000)	17	(3700)				
12	8	5	2	(100)	12	(1720)				
13	8	5	2	(100)	8	(920)	12	(1480)	15	(2700)
14	8	5	2	(100)	8	(920)	14	(2000)	17	(3700)
23	8	5	8	(770)	15	(2700)				
24	8	5	8	(770)	12	(1720)	14	(2000)	17	(3700)
34	8	5	12	(1480)	17	(3700)				
123	8	5	2	(100)	15	(2700)				
124	8	5	2	(100)	12	(1720)	14	(2000)	17	(3700)
134	8	5	2	(100)	8	(920)	12	(1480)	17	(3700)
234	8	5	8	(770)	17	(3700)				

Table 9: Corresponding feature structure for the first series of experiments

[a] cb = critical bands

[b] cb2 = critical bands for the second set of frequencies for non–neighbouring subbands combinations.

[14] For programme–source, see: /homes/hagen/STRUT/src/strut-1.08e/lpc-cepstrum-several-bands.cc

C Description of J–Rasta–PLP Features for the Different Subband Combinations

In the following Table 10 the parameters for the calculation of the J–Rasta–PLP Subband Features are summarized, which were used by the STRUT programme 'lpc-cepstrum-bands' and 'lpc-cepstrum-several-bands'[15]. The PLP–feautres were J–Rasta filtered before processing into the different subbands.

bands	lpc–cepstrum–bands									
	coeff.	lpc order	1^{st} cb[a]	(Hz)	last cb	(Hz)	1^{st} cb2[b]	(Hz)	last cb2	(Hz)
1	8	5	2	(100)	8	(920)				
2	8	2	8	(770)	11	(1480)				
3	8	2	11	(1270)	15	(2700)				
4	8	2	14	(2000)	16	(3150)				
12	8	5	2	(100)	11	(1480)				
13	8	5	2	(100)	8	(920)	12	(1480)	14	(2320)
14	8	5	2	(100)	8	(920)	14	(2000)	16	(3150)
23	8	5	8	(770)	14	(2320)				
24	8	5	8	(770)	11	(1480)	14	(2000)	16	(3150)
34	8	3	12	(1480)	16	(3150)				
123	8	5	2	(100)	14	(2320)				
124	8	5	2	(100)	11	(1480)	14	(2000)	16	(3150)
134	8	5	2	(100)	8	(920)	12	(1480)	16	(3150)
234	8	5	8	(770)	16	(3150)				

Table 10: Corresponding feature structure for the first series of experiments

[a] cb = critical bands

[b] cb2 = critical bands for the second set of frequencies for the non–neighbouring subbands.

[15] For programme–source, see: /homes/hagen/STRUT/src/strut-1.08e/lpc-cepstrum-several-bands.cc

D Single Results of the 15 trained MLPs

In the following Tables 11 to 16 the results of the 15, single MLPs are summarized. They were separately tested to see their respective performance for both the PLP-Rasta features and the J-Rasta-PLP features. Tests were carried out on clean speech as well as speech corrupted by car noise of -30 to +30 dB SNR.

The Figures 11 and 12 illustrate the word error rates of each of the 15 MLPs versus changing SNR values. In Figures 11 and 12, the black graphs show the MLPs consisting of one subband, the blue graphs the MLPs of two subbands, the red graphs the MLPs existing of three subbands and the fullband MLP is shown in green. The best combination of MLPs that could be achieved in the framework of this article is shon in cyan.

For the PLP-features, the results seem to depend more on the respective subbands included in the combination than on the number of included subbands. The combinations including subband 4 (MLP 124, MLP 134, MLP 4) have the highest word error rates. These follow MLP 3 and MLP 34, then the word error rates of the one-band MLPs one and two. Among the best MLPs for the PLP-features are those consisting of bands 1, 2 and 3 such as MLP 123. The fullband hybrid system has the lowest error rate only between 10 db and 45 db SNR. The best combination system achieves clearly smaller error rates for the nosiy data than the fullband system and almost the best rates for clean data as well.

For the MLPs trained on the J-Rasta-PLP features, it can clearly be seen that the MLPs which were trained on one band only have almost always the highest word error rates (black). They are followed by the MLPs trained on two subbands (blue). After these come the error rates of the MLPs trained on three subbands (red). The fullband hybrid system has the lowest error rates for low SNR values and the second best for high SNRs, what can be seen at the green graph. The best combination system (cyan) is better for the high SNR values, but less efficient for noisy data.

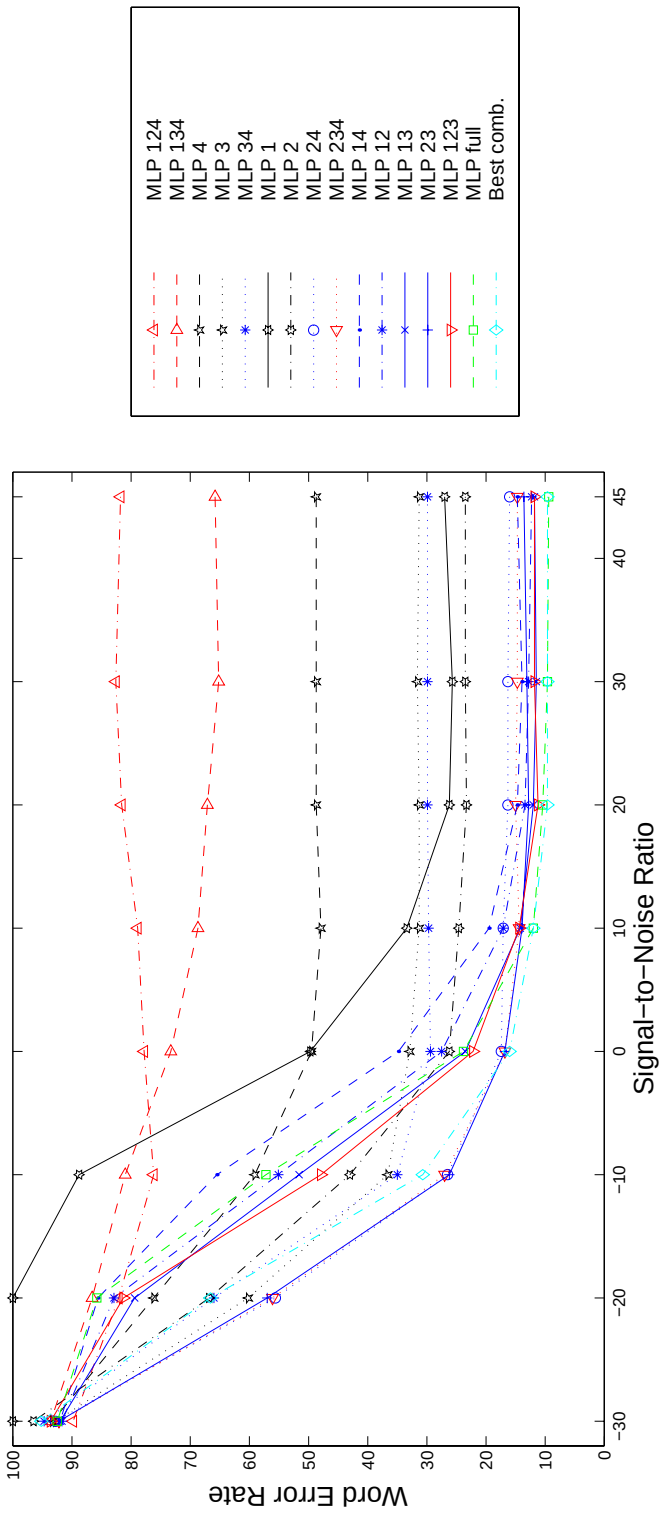


Figure 11: Illustration of the single trained MLPs, incorporating the PLP-Rasta Features

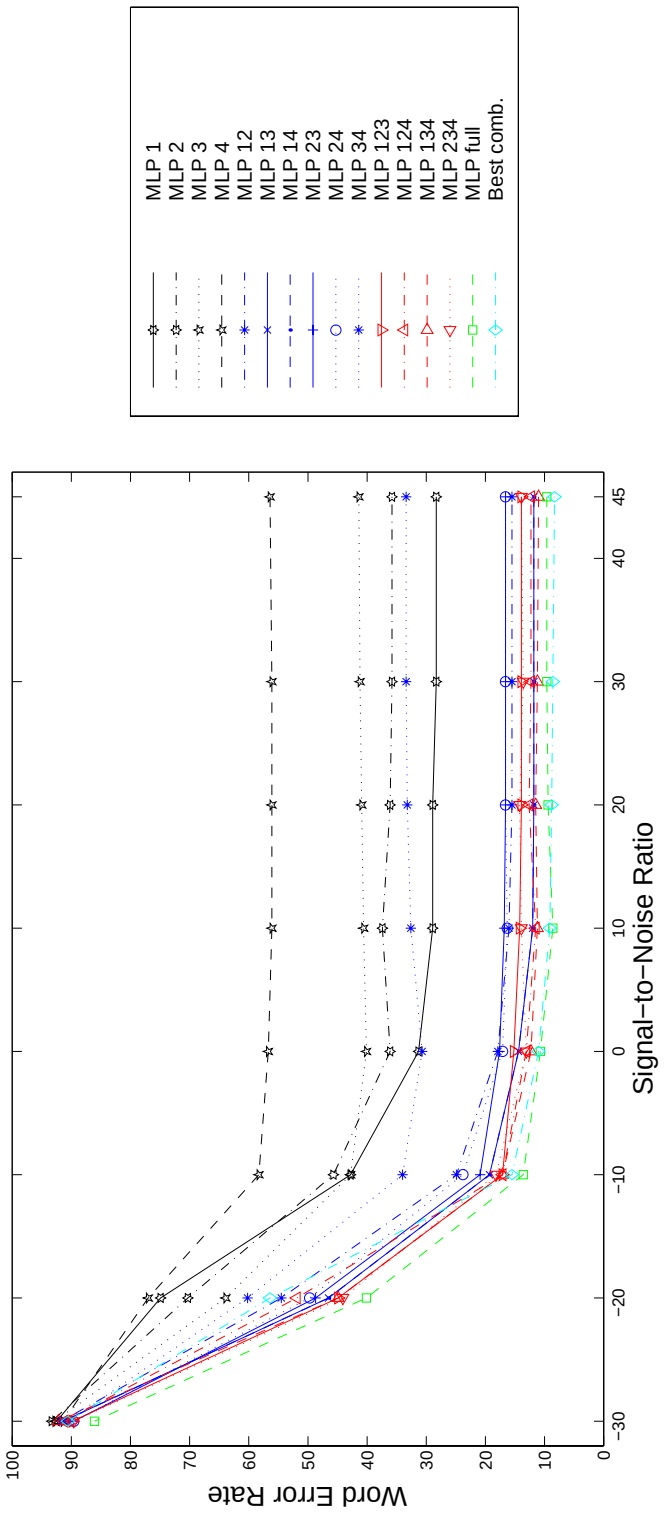


Figure 12: Illustration of the single trained MLPs, incorporationg the Jah-Rasta Features

MLP	Insertion	Deletion	Substitution	Word Err	Sentence Err
1	5.6	1.9	19.5	**27.0**	59.0
+30	4.8	2.1	18.7	**25.7**	59.0
+20	4.8	2.4	19.0	**26.2**	60.0
+10	3.5	7.5	22.5	**33.4**	70.0
0	1.6	25.7	22.5	**49.7**	87.0
-10	0.0	77.5	11.2	**88.8**	99.0
-20	0.0	100.0	0.0	**100.0**	100.0
-30	0.0	100.0	0.0	**100.0**	100.0
2	5.9	4.0	13.6	**23.5**	49.0
+30	5.9	4.0	13.6	**23.5**	48.0
+20	5.1	4.0	14.2	**23.3**	48.0
+10	4.0	4.5	16.0	**24.6**	55.0
0	2.4	8.8	15.0	**26.2**	58.0
-10	0.3	20.3	22.5	**43.0**	78.0
-20	0.0	48.9	17.9	**66.8**	95.0
-30	0.0	93.3	3.2	**96.5**	99.0
3	4.5	7.5	19.3	**31.3**	59.0
+30	4.5	7.5	19.5	**31.6**	59.0
+20	4.5	7.5	19.3	**31.3**	59.0
+10	4.0	7.2	20.1	**31.3**	58.0
0	3.5	9.1	20.3	**32.9**	61.0
-10	2.9	13.4	20.3	**36.6**	67.0
-20	2.1	33.2	24.9	**60.2**	89.0
-30	0.3	82.4	10.2	**92.8**	97.0
4	7.2	9.4	32.1	**48.7**	82.0
+30	7.5	9.4	31.8	**48.7**	82.0
+20	7.0	9.4	32.4	**48.7**	81.0
+10	6.1	9.6	32.1	**47.9**	80.0
0	5.9	10.7	32.9	**49.5**	80.0
-10	7.0	18.4	33.7	**59.1**	84.0
-20	2.4	41.4	32.4	**76.2**	94.0
-30	0.8	85.0	7.5	**93.3**	97.0
12	3.2	1.9	7.2	**12.3**	30.0
+30	3.5	1.6	7.8	**12.8**	32.0
+20	4.0	1.6	7.8	**13.4**	33.0
+10	5.1	1.9	10.2	**17.1**	39.0
0	8.6	2.1	16.8	**27.5**	57.0
-10	11.2	4.3	39.6	**55.1**	86.0
-20	8.8	19.5	54.5	**82.9**	97.0
-30	3.2	51.6	36.9	**91.7**	98.0
13	2.1	1.9	7.8	**11.8**	29.0
+30	2.1	1.6	7.8	**11.5**	29.0
+20	2.4	1.6	8.0	**12.0**	29.0
+10	4.0	1.3	8.8	**14.2**	34.0
0	6.4	2.4	14.7	**23.5**	50.0
-10	12.6	5.3	33.7	**51.6**	82.0
-20	7.5	17.4	54.5	**79.4**	95.0
-30	1.9	56.1	34.0	**92.0**	98.0

Table 11: PLP-Rasta Features

MLP	Insertion	Deletion	Substitution	Word Err	Sentence Err
14	4.0	1.3	9.4	**14.7**	35.0
+30	3.7	1.3	8.8	**13.9**	33.0
+20	4.3	1.3	9.1	**14.7**	35.0
+10	5.6	1.1	12.8	**19.5**	43.0
0	12.6	2.4	19.8	**34.8**	64.0
-10	18.4	4.5	42.5	**65.5**	88.0
-20	7.0	19.8	58.8	**85.6**	98.0
-30	0.3	62.0	29.7	**92.0**	99.0
23	3.5	2.1	8.0	**13.6**	31.0
+30	3.7	1.9	7.5	**13.1**	31.0
+20	3.2	1.9	7.8	**12.8**	31.0
+10	3.2	2.4	8.3	**13.9**	33.0
0	3.2	2.1	11.5	**16.8**	42.0
-10	1.9	7.5	16.8	**26.2**	58.0
-20	1.6	31.3	24.1	**57.0**	90.0
-30	0.0	86.4	5.9	**92.2**	98.0
24	4.5	2.4	9.1	**16.0**	36.0
+30	4.8	2.4	9.1	**16.3**	37.0
+20	4.5	2.4	9.4	**16.3**	38.0
+10	4.0	2.7	10.4	**17.1**	39.0
0	2.7	2.9	11.8	**17.4**	41.0
-10	2.4	7.0	17.1	**26.5**	60.0
-20	0.3	33.7	21.7	**55.6**	86.0
-30	0.0	87.7	5.1	**92.8**	98.0
34	5.9	6.4	17.6	**29.9**	57.0
+30	5.9	6.4	17.6	**29.9**	57.0
+20	5.9	6.4	17.6	**29.9**	57.0
+10	5.1	6.4	18.2	**29.7**	57.0
0	4.0	6.7	18.7	**29.4**	57.0
-10	3.7	12.0	19.3	**35.0**	65.0
-20	5.3	35.8	24.9	**66.0**	91.0
-30	2.4	77.8	14.4	**94.7**	97.0
123	3.5	1.9	6.4	**11.8**	29.0
+30	3.5	1.9	6.4	**11.8**	29.0
+20	3.2	1.3	6.7	**11.2**	28.0
+10	4.5	1.1	8.8	**14.4**	32.0
0	7.8	2.1	12.3	**22.2**	44.0
-10	12.0	4.3	31.6	**47.9**	80.0
-20	9.6	17.9	53.7	**81.3**	95.0
-30	3.2	49.5	40.9	**93.6**	98.0
124	14.7	11.0	56.1	**81.8**	99.0
+30	15.2	11.0	56.4	**82.**6	99.0
+20	14.4	12.6	54.5	**81.**6	99.0
+10	13.1	13.9	51.9	**78.**9	98.0
0	12.8	17.9	47.1	**77.**8	97.0
-10	13.4	18.2	44.7	**76.**2	97.0
-20	8.0	34.2	39.8	**82.1**	97.0
-30	1.1	69.3	19.5	**89.8**	98.0

Table 12: PLP-Rasta Features — cont.

MLP	Insertion	Deletion	Substitution	Word Err	Sentence Err
134	6.1	15.5	44.1	**65.8**	94.0
+30	5.9	16.3	43.0	**65.2**	94.0
+20	6.4	17.1	43.6	**67.1**	94.0
+10	7.0	19.3	42.5	**68.7**	93.0
0	10.4	16.8	46.0	**73.3**	95.0
-10	11.2	16.3	53.5	**81.0**	94.0
-20	10.7	20.3	55.6	**86.6**	98.0
-30	5.1	41.4	47.1	**93.6**	99.0
234	4.0	2.1	8.6	**14.7**	33.0
+30	4.0	2.1	8.6	**14.7**	33.0
+20	4.0	2.1	8.8	**15.0**	34.0
+10	3.5	1.9	9.1	**14.4**	34.0
0	3.2	3.2	10.4	**16.8**	40.0
-10	1.9	7.5	17.6	**27.0**	58.0
-20	1.1	33.2	21.9	**56.1**	88.0
-30	0.0	87.2	5.1	**92.2**	97.0
full	2.7	1.1	5.6	**9.4**	23.0
+30	2.7	1.1	5.6	**9.4**	23.0
+20	2.7	1.1	6.7	**10.4**	27.0
+10	3.5	1.3	7.8	**12.6**	30.0
0	6.7	1.6	18.7	**27.0**	57.0
-10	11.2	6.4	46.3	**63.9**	89.0
-20	4.5	54.8	31.8	**91.2**	99.0
-30	4.0	43.6	48.7	**96.3**	99.0

Table 13: PLP-Rasta Features — cont.

MLP	Insertion	Deletion	Substitution	Word Err	Sentence Err
1	6.7	2.1	19.5	**28.3**	61.0
+30	6.7	2.1	19.5	**28.3**	61.0
+20	6.7	2.1	20.1	**28.9**	62.0
+10	6.7	1.6	20.6	**28.9**	62.0
0	5.6	2.1	23.5	**31.3**	66.0
-10	7.5	5.3	29.9	**42.8**	79.0
-20	5.9	25.7	43.3	**74.9**	99.0
-30	1.1	72.2	19.3	**92.5**	100.0
2	5.3	4.3	26.2	**35.8**	70.0
+30	5.3	4.3	26.2	**35.8**	70.0
+20	5.6	4.3	26.2	**36.1**	70.0
+10	5.6	4.8	27.0	**37.4**	70.0
0	5.6	4.3	26.2	**36.1**	70.0
-10	1.9	10.4	33.4	**45.7**	83.0
-20	2.1	27.8	40.4	**70.3**	96.0
-30	1.3	65.0	27.0	**93.3**	100.0
3	9.1	6.4	25.9	**41.4**	70.0
+30	8.8	6.4	25.9	**41.2**	70.0
+20	8.8	6.4	25.7	**40.9**	69.0
+10	8.6	6.7	25.4	**40.6**	69.0
0	8.0	7.0	25.1	**40.1**	69.0
-10	7.0	10.7	25.1	**42.8**	71.0
-20	2.9	27.0	34.0	**63.9**	86.0
-30	1.3	74.1	17.4	**92.8**	98.0
4	8.6	7.8	40.1	**56.4**	87.0
+30	8.6	7.5	40.1	**56.1**	87.0
+20	8.6	7.5	40.1	**56.1**	87.0
+10	8.3	8.0	39.8	**56.1**	87.0
0	9.1	9.1	38.5	**56.7**	87.0
-10	7.0	12.8	38.5	**58.3**	90.0
-20	6.1	32.6	38.2	**77.0**	93.0
-30	1.3	68.4	21.7	**91.4**	99.0
12	4.8	1.1	9.6	**15.5**	33.0
+30	4.8	1.1	9.6	**15.5**	33.0
+20	4.8	1.1	9.6	**15.5**	33.0
+10	4.5	1.1	10.4	**16.0**	34.0
0	5.6	1.3	11.0	**17.9**	41.0
-10	4.3	2.7	17.9	**24.9**	57.0
-20	2.1	23.5	28.9	**54.5**	91.0
-30	0.0	80.2	11.5	**91.7**	99.0
13	3.7	1.1	7.0	**11.8**	32.0
+30	3.7	1.1	7.0	**11.8**	32.0
+20	3.7	1.1	7.0	**11.8**	32.0
+10	3.7	1.1	7.2	**12.0**	32.0
0	4.5	0.8	9.1	**14.4**	38.0
-10	5.6	1.9	11.8	**19.3**	47.0
-20	3.2	19.0	24.3	**46.5**	82.0
-30	0.3	79.7	11.5	**91.4**	100.0

Table 14: Jah-Rasta Features

MLP	Insertion	Deletion	Substitution	Word Err	Sentence Err
14	5.1	1.6	9.1	**15.8**	36.0
+30	4.8	1.6	9.4	**15.8**	36.0
+20	4.8	1.6	9.1	**15.5**	35.0
+10	4.5	1.3	9.6	**15.5**	35.0
0	5.6	1.3	10.2	**17.1**	39.0
-10	5.6	3.2	15.2	**24.1**	54.0
-20	4.3	20.9	32.6	**57.8**	90.0
-30	0.3	78.3	14.4	**93.0**	100.0
23	4.3	2.9	9.4	**16.6**	38.0
+30	4.3	2.9	9.4	**16.6**	38.0
+20	4.3	2.9	9.4	**16.6**	38.0
+10	4.3	2.9	9.6	**16.8**	39.0
0	3.7	2.7	11.2	**17.6**	41.0
-10	2.1	5.3	13.4	**20.9**	45.0
-20	1.3	23.8	23.5	**48.7**	79.0
-30	1.1	75.1	13.4	**89.6**	97.0
24	3.5	1.9	11.2	**16.6**	39.0
+30	3.5	1.9	11.2	**16.6**	39.0
+20	3.5	1.9	11.2	**16.6**	39.0
+10	3.2	1.6	11.5	**16.3**	40.0
0	3.5	2.4	11.2	**17.1**	41.0
-10	2.4	4.5	16.8	**23.8**	51.0
-20	1.3	25.7	22.7	**49.7**	80.0
-30	0.5	74.3	14.7	**89.6**	98.0
34	7.0	6.1	20.3	**33.4**	63.0
+30	7.0	6.1	20.3	**33.4**	63.0
+20	6.7	5.9	20.6	**33.2**	63.0
+10	5.9	5.6	21.1	**32.6**	64.0
0	5.1	6.7	19.0	**30.7**	62.0
-10	4.8	10.2	19.0	**34.0**	62.0
-20	3.7	26.5	29.9	**60.2**	83.0
-30	1.3	71.9	17.1	**90.4**	97.0
123	4.5	1.3	8.0	**13.9**	30.0
+30	4.5	1.3	8.0	**13.9**	30.0
+20	4.5	1.3	8.0	**13.9**	30.0
+10	4.8	1.3	8.0	**14.2**	30.0
0	5.1	1.3	8.8	**15.2**	34.0
-10	4.3	2.7	10.2	**17.1**	41.0
-20	2.9	19.5	22.5	**44.9**	84.0
-30	0.5	79.4	9.9	**89.8**	99.0
124	3.5	1.3	7.5	**12.3**	28.0
+30	3.5	1.3	7.5	**12.3**	28.0
+20	3.7	1.3	7.5	**12.6**	28.0
+10	3.2	1.1	7.2	**11.5**	25.0
0	4.0	1.6	7.5	**13.1**	29.0
-10	3.7	2.4	11.0	**17.1**	43.0
-20	2.9	20.3	28.6	**51.9**	85.0
-30	0.3	81.3	9.6	**91.2**	99.0
134	2.9	1.1	7.0	**11.0**	26.0
+30	2.9	1.3	7.0	**11.2**	27.0
+20	2.9	1.3	7.2	**11.5**	28.0
+10	3.2	1.1	7.0	**11.2**	28.0
0	3.2	1.1	8.0	**12.3**	31.0
-10	3.7	2.1	11.2	**17.1**	42.0
-20	2.9	17.1	24.9	**44.9**	81.0
-30	0.3	79.9	12.0	**92.2**	100.0

Table 15: Jah-Rasta Features — cont.

MLP	Insertion	Deletion	Substitution	Word Err	Sentence Err
234	2.7	2.1	9.1	**13.9**	35.0
+30	2.4	2.1	9.1	**13.6**	35.0
+20	2.7	2.4	9.1	**14.2**	36.0
+10	2.9	2.1	8.8	**13.9**	34.0
0	2.4	2.4	8.3	**13.1**	33.0
-10	1.6	5.6	11.0	**18.2**	43.0
-20	1.1	24.3	18.7	**44.1**	75.0
-30	0.8	75.9	12.8	**89.6**	96.0
full	2.9	1.6	5.1	**9.6**	23.0
+30	2.9	1.6	5.1	**9.6**	23.0
+20	2.7	1.3	5.3	**9.4**	23.0
+10	2.1	1.1	5.3	**8.6**	22.0
0	2.9	1.1	6.7	**10.7**	27.0
-10	2.7	1.1	9.9	**13.6**	37.0
-20	2.9	15.2	21.9	**40.1**	81.0
-30	0.3	73.5	12.3	**86.1**	97.0

Table 16: Jah-Rasta Features — cont.

References

[BD96] H. Bourlard and S. Dupont. A new asr approach based on independent processing and recombination of partial frequency bands. *Int. Conf. on Spoken Language Processing*, pages 426–429, 1996.

[BD97] H. Bourlard and S. Dupont. Subband–based speech recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, pages 1251–1254, 1997.

[BD98] H. Bourlard and S. Dupont. Personal communication, 1998.

[BDR96] H. Bourlard, S. Dupont, and C. Ris. Multi–stream speech recognition. *IDIAP Research Report*, pages 1–13, 1996.

[Bis95] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

[BRA95] E. Bocchieri, G. Riccardi, and J. Anantharaman. The 1994 at&t atis chronus recognizer. *ARPA Spoken Language Systems Technology Workshop*, pages 265–268, 1995.

[CNLD95] R.A. Cole, M. Noel, T. Lander, and T. Durham. New telephone speech corpora at cslu. *Proc. European Conf. on Speech Communication and Technology*, 1:821–824, 1995.

[DB96] S. Dupont and H. Bourlard. Multiband approach for speech recognition. *Proc. of ProRISC/IEEE Workshop on Circuits, Systems and Signal Processing, Mierlo, The Netherlands*, pages 113–118, 1996.

[DBR97] S. Dupont, H. Bourlard, and C. Ris. Robust speech recognition based on multi–stream features. *Proc. ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-à-Mousson, France*, pages 95–98, 1997.

[Her90] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, April 1990.

[Her98] H. Hermansky. Should recognizers have ears? *Speech Communication*, 25:3–27, 1998.

[Hir93] H.G. Hirsch. Estimation of noise spectrum and its application to snr-estimation and speech enhancement. *ICSI Technical Report*, pages 1–32, 1993.

[HM94] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, October 1994.

[HMBK92] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn. Rasta–plp speech analysis technique. *IEEE Trans. on Signal Processing*, 1:121–124, 1992.

[HTP96] H. Hermansky, S. Tibrewala, and M. Pavel. Towards asr on partially corrupted speech. *Int. Conf. on Spoken Language Processing*, pages 462–465, 1996.

[LM94] T. Lander and S.T. Metzler. The cslu labeling guide. *CSLU, Oregon*, 1994.

[MCG98] A. C. Morris, M. P. Cooke, and P. D. Green. Some solutions to the missing features problem in data classification, with application to noise robust asr. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pages 737–740, 1998.

[OBP98] S. Okawa, E. Bocchieri, and A. Potamianos. Multi–band speech recognition in noisy environment. *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1998.

[VSTJ92] A. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones. The noiesex–92 study on the effect of additive noise on automatic speech recognition. *Technical Report, DRA Speech Research Unit*, 1992.