DIFFERENT WEIGHTING SCHEMES IN THE FULL COMBINATION SUBBANDS APPROACH FOR NOISE ROBUST ASR[1]

Astrid Hagen, Andrew Morris, Hervé Bourlard[2]

Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP)
P.O.Box 592, rue du Simplon 4, 1920 Martigny Switzerland
e-mail: hagen,morris,bourlard@idiap.ch

ABSTRACT

In this paper, we present and investigate a new method for subband-based Automatic Speech Recognition (ASR) which approximates the ideal 'full combination' approach which is itself often not practical to realize. The 'full combination' approach consists of explicitly considering all possible combinations of subbands [6] avoiding the usually necessary independence assumption, which would limit the potential of subband-based ASR.

We show how this ideal approach can be effectuated by a nonlinear combination function which constitutes the full-band posterior probabilities decomposed into a weighted sum of posterior probabilities from Artificial Neural Network (ANN) experts. This involves training of one expert for each possible subband combination. To limit such extensive training, we have found that it is possible to achieve comparable results by estimating the subband posterios for each combination as a function of the posteriors from the individual subbands alone [4, 8].

The theoretical foundation of our solution to the ideal 'full combination' approach with the nonlinear combination function and its approximation are presented. The weights, which represent the relative utility for recognition of each subband combination, are very important for this technique and possible schemes for their estimation will be proposed. They have been tested and compared in the framework of HMM/ANN-Hybrid systems on clean and noise-added data.

1. INTRODUCTION

Noise interfering with the speech signal can often be found to occur in restricted frequency bands only. As results from 'Missing Data' (MD) theory have shown [7, 9, 5] recognition can be strongly improved when these noisy subbands can be detected and ignored. Employing subband-based recognizers which can disregard these noisy subbands can lead to a considerable improvement in recognition rate in noise. However, subband based ASR has so far been limited by the necessary assumption of subband independence: as the subband paradigm requires that we work in the spectral domain, the subbands are necessarily correlated. If we orthogonalize the components of the frequency band, we are no longer in the spectral domain and have already spread the noise; thus, in order to process frequency subbands separately, the independence assumption is made.

This independence assumption can only be avoided if a separate recognizer is trained on each possible combination of subbands [5], but this still leaves us with the problem of how to combine the outputs from all recognizers. In this paper we show how the fullband posterior (phoneme) probabilities can be decomposed into a weighted sum of the posteriors from the subband recognizers alone – the *Full Combination (FC) method*. Moreover, we have found that it is possible to achieve comparable results by estimating the posteriors of all the possible subband combinations as a function of the posteriors from the individual sub-bands alone. As opposed to other sub-band-based approaches, the proposed solution is more mathematically consistent and also allows us to relax some of the sub-band independence assumption.

The weighting factors in the decomposition function denote the relative reliability of the individual sub-band combinations. We will see how the weighting factors can be either chosen as constant or approximated from the segmental signal-to-noise ratio (SNR) estimates for the individual sub-bands.

The experiments for the Full Combination method and its approximation were carried out on the Numbers95 database [2] with car noise from the Noisex92 databaseadded at different SNR values.

2. THEORETICAL FRAMEWORK

Most of the sub-band based ASR approaches developed previously consist in splitting the frequency range into several,

[1] To appear in Proceedings of the ESCA Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, Mai 1999

[2] Also Professor at the Swiss Federal Institute of Technology, Lausanne, CH.

independently processed bands, and in feeding the resulting sub-band features into independent recognizers. The sub-band posterior (phoneme) probabilities are then combined later in the recognition process at some segmental level. Ideally, this approach should consider all possible sub-band combinations and select the best one, as confirmed by the experiments reported in [6]. However, since firstly it is not always feasible to consider all possible combinations and secondly it is very difficult to automatically select the best sub-band combination, most of the sub-band approaches use simple combination schemes of a few disjoint sub-bands, assuming that the frequency bands are independent and that the noise is limited to one of these bands [1, 3].

2.1. Full Combination (FC) of phoneme posterior probabilities

We now generalize this sub-band approach by considering all possible sub-band combinations and show how to actually combine the evidence from all sub-band subsets. As illustrated in Figure 1 for the case of two sub-bands x_1 and x_2, we would like to estimate and combine the (posterior) probabilities from all possible sub-band 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 sub-bands is reliable).

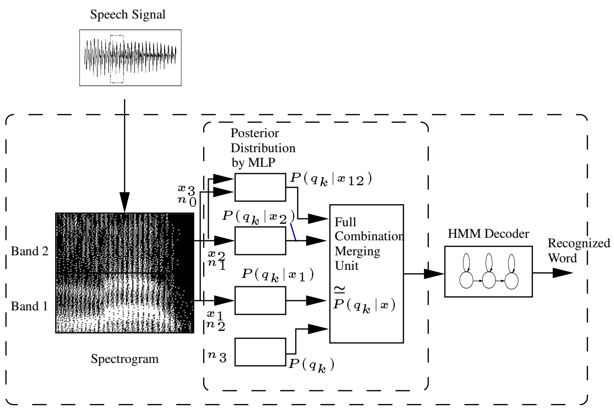

Figure 1: Illustration of the full combination approach on two sub-bands.

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 sub-bands 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 value. If we have acoustic vectors of dimension d (or d sub-bands), we have $C = 2^d$ possible 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:

$$P(q_k|x) \quad = \quad \sum_{i=1}^{C} P(q_k|x, n_i) P(n_i|x) \qquad (1)$$

We now consider how to estimate each of the terms in (1). 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 [7, 9, 5] have shown that, when the position of highly inaccurate or missing data is known, recognition can be strongly improved simply by ignoring this data. In our case this corresponds to simply equating the joint information (x, n_i) in (1), 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}[1]. Consequently (1) becomes:

$$P(q_k|x) \; = \sum_{i=1}^{C} P(q_k|x_{c_i}) P(n_i|x) \qquad (2)$$

$P(n_i|x_{c_i})$ 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:

- How to compute the sum in (2) over all the C possible combinations? A direct approach would be, for each term of the sum, i.e. for each possible combination of sub-bands, 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 sub-bands, 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 2.2, we will propose an approximation that allows us to avoid training for every sub-band combination, and where every term $P(q_k|x_{c_i})$ required in (2) is estimated on the basis of a minimal set of neural networks, typically the single sub-band neural networks.

- How to estimate the weighting factors $P(n_i|x)$ in (2)? These weights, which represent the relative utility for recognition of each sub-band combination x_{c_i}, are very important for this technique and different schemes for their estimation will be shown in Section2.3.

[1] c_i will denote the set of indices for sub-bands in subset i, so that $x_{c_i} = x_j : j \in c_i$

2.2. Approximation to the FC Method

Computation of (2) requires the posteriors $P(q_k|x_{c_i})$ for each of K states from the neural networks trained for all $C = 2^d$ different possible combinations of d sub-bands. In the experiments which follow the number of sub-bands was limited to 4 (resulting in 16 neural networks) so this was feasible, but this approach rapidly becomes impractical as the number of sub-bands increases.

This problem would clearly be solved if the C sub-band combination posteriors $P(q_k|x_{c_i})$ for each state q_k could be expressed in terms of the single sub-band posterior probabilities $P(q_k|x_j)$, $j = (1, ..., d)$, alone. While it is not possible to obtain the exact combination posteriors in this way, it is possible to approximate combination posteriors from single sub-band posteriors, without assuming full sub-band independence but only conditional independence on q_k, by the following procedure [4, 8]:

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

$$\hat{P}(q_k|x_{c_i}) = \frac{\bar{P}_{ki}}{\sum_{l=1}^{K} \bar{P}_{li}} \tag{3}$$

This is the equation used in the experiments to the approximation of the FC approach.

2.3. Different Weighting Schemes

Let's now interpret the factors $P(n_i|x)$ in (2) 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 sub-band in sub-band combination i is clean, and every other sub-band (the components disregarded in the computation of $P(q_k|x_{c_i})$) is noisy. On the assumption that the presence of noise in each sub-band is independent, this probability $P(n_i|x)$ can be approximated as the product of the probabilities for each sub-band in subset i being clean, and each remaining sub-band being noisy [2].

$$\hat{P}(n_i|x) \simeq \prod_{j \in c_i} P(x_j \; clean) \cdot \prod_{k \notin c_i} P(x_k \; noisy) \tag{4}$$

The factors on the right hand side of (4) were estimated as a linear function of the *SNR estimator* working in each of the d frequency bands. The experiments show that even when using *a-priori SNR* values, the probabilities estimated in this way lead to less good results than using equal weights for each combination. This shows that our estimate for $P(x_i \; clean)$ from the SNR estimate of band j is presently far from optimal.

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

3. EXPERIMENTS

Experiments were carried out in the framework of HMM-MLP-Hybrid systems using two separate sets of acoustic features. The first set are the PLP (Perceptual Linear Prediction) features to evaluate the new approach on well-known and well-performing features. The second set are features that are proven to be more noise robust: J-Rasta-PLP features. The 16 Full Combination MLPs, as well as the reference fullband MLP, were trained on 9 frames of contextual input, with one hidden layer of 1000 hidden units and an output layer with 33 units, one for each phoneme. The four sub-bands comprise the frequency ranges of [17-949 Hz], [707-1632 Hz], [1506-2709 Hz] and [2122-3769 Hz].

Tests were run on the first 100 utterances from Numbers95's test set. For the experiments with noise corrupted data, car noise from the Noisex92 database was added to the clean speech at SNR rates of -10, 0, 10 and 20 dB.

In the following section, we present the experiments and results for the full combination (Equation (2)) approach and its approximation (Equation (3)) incorporating equal and SNR-based weighting schemes.

System	Signal-To-Noise Ratio				clean
	-10	0	10	20	45
Fullband	60.2	25.4	14.7	11.0	10.4
Early Sub. equ.w.	39.3	24.1	17.6	16.8	16.3
FC equal w.	35.0	15.8	11.5	9.6	9.6
FC SNR w.	34.5	17.4	11.2	8.6	8.0
Approx. equ.	34.8	20.3	11.8	12.3	12.8
Approx. SNR	34.5	20.1	13.9	13.1	13.4

Table 1: Word Error Rate for PLP-Features with Full-Combination (FC) and its Approximation on Car Noise

First experiments were carried out with the PLP-Features. Results can be seen in Table 1 and Figure 2 for car-noise added speech. For comparison also the "early sub-band approach" [3] was tested which recombines just the four MLPs, trained on one sub-band each, in a (weighted) sum without further approximating missing combinations of sub-bands (line 2). The FC method consisting of the 16 trained MLPs (lines 3-4) and its approximation by 4 MLPs (lines 5-6) were both tested with equal weights, estimated SNR-based weights (cf. Equation 4) and a-priori calculated SNR-based weights as pointed out above. They are furthermore compared to the MLP trained on the full frequency domain (line 1). Results show that the FC method improved recognition rates on clean and noisy data for all weighting schemes as compared to the fullband or the early sub-band approaches [3]. The estimated SNR-weights or even the a-priori SNR weights, in the FC method though did not improve recognition rates as over equal weighting. The approximation to the FC method by the 4 MLPs also resulted in higher recognition rates than the fullband or the early sub-band systems but did not achieve the same results as the FC method itself.

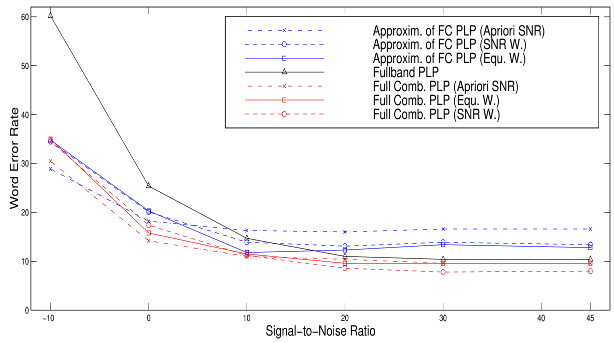

Figure 2: Illustration of the full combination approach and its approximation (with different weighting schemes) as compared to the fullband hybrid with PLP features on car noise

We then ran the same set of experiments on the J-Rasta-PLP features. Using these features already resulted in higher recognition rates, especially for noise corrupted data. Therefore, it was very hard to further improve the results by employing the FC method and its approximation. Results are illustrated in Figure 3. This time, no significant improvement could be achieved with the FC method. For the J-Rasta-PLP features, the approximation by the 4 MLPs could not approximate the results of the FC system. Again, a-priori weighting resulted in no improvement.

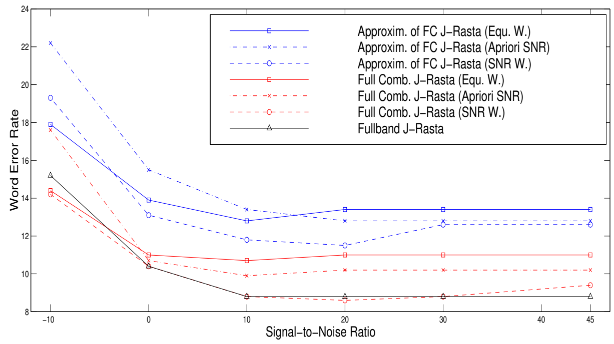

Figure 3: Illustration of the full combination approach and its approximation (with different weighting schemes) as compared to the fullband hybrid with J-Rasta-PLP features on car noise.

4. CONCLUSIONS

The proposed 'full combination' method for sub-band-based ASR seems to have a good potential to improve recognition rates on noise corrupted data, but is highly sensitive to the chosen features and weighting strategy. The results from the different weighting schemes show that the value of the output from each sub-band combination MLP may be dependent not only on the noise level in the sub-bands concerned, but also on the inherent utility of the information normally found in the sub-bands concerned for speech recognition. We therefore want to look for noise independent weighting schemes for each of the sub-band combination MLPs which possibly also consider each recognition unit (phoneme) separately. A *Least Mean Square Error (LMSE) Calculation* or *Expectation-Maximazation Training* of the weights could result in such noise independent, combination- and phoneme-specific weights which we want to investigate in the future.

Acknowledgments:

This work is supported by the Swiss Federal Office for Education and Science (FOES) through the European TMR SPHEAR project, and by the Swiss National Science Fund project MULTICHAN.

References

[1] H. Bourlard, S. Dupont, and C. Ris. Multi–stream speech recognition. IDIAP-RR 07, IDIAP, 1996.

[2] 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.

[3] 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.

[4] A. Hagen, A. Morris, and H. Bourlard. Subband-based speech recognition in noisy conditions: The full combination approach. IDIAP-RR 15, IDIAP, 1998.

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

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

[7] R. P. Lippmann and B. A. Carlson. Using missing feature theory to actively select features for robust speech regonition with interruptions filtering and noise. *Proc. European Conf. on Speech Communication and Technology*, pages 37–40, 1997.

[8] A. Morris, A. Hagen, and H. Bourlard. The full combination sub-bands approach to noise robust hmm/ann-based asr. *Proc. European Conf. on Speech Communication and Technology*, 1999. To Appear.

[9] 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.