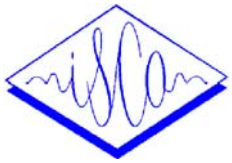


6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

TOWARDS A GLOBAL OPTIMIZATION SCHEME FOR MULTI-BAND SPEECH RECOGNITION

Cerisara Christophe, Haton Jean-Paul, Fohr Dominique
LORIA / INRIA Lorraine,
Campus Scientifique, B.P. 239
54506 Vandoeuvre-les-Nancy Cedex, FRANCE
{cerisara, jph, fohr}@loria.fr

ABSTRACT

In this paper, we deal with a new method to globally optimize a Multi-Band Speech Recognition (MBSR) system. We have tested our algorithm with the TIMIT database and obtained a significant improvement in the accuracy over a basic HMM system for clean speech. The goal of this work is not to prove the effectiveness of MBSR, what has yet been done, but to improve the training scheme by introducing a global optimization procedure. A consequence of this method is that the models are no longer phone-models, but define new classes in the phonetic space which might better model the acoustic information carried by each band.

Keywords: Multi-Band, Speech Recognition, Global Training.

1. INTRODUCTION

The Multi-Band paradigm has already been precisely described in several previous papers [Cerisara98] [Bourlard97] [Tibrewala97]. To summarize the basic principle, MBSR consists of filtering the signal into several frequency bands on which are applied independent phone recognizers. The answers of these recognizers are merged together in the final stage of the system. The main motivations for Multi-Band paradigm are first, its robustness to frequency-limited noise, and then the fact that the information which characterizes a phoneme is often contained in a limited region of the spectrum. The dual of this idea is that there is *not enough* information in a single frequency band to identify all the phonemes, i.e., phone models are not adapted to sub-band recognition. This is why the idea of using new classes in each band, which are intermediate between phones and broad phonetic classes, has recently emerged. There are three major advantages to use such classes:

- As there is less information in each band, there are also less classes to recognize. That means that more training examples are used for each class, and that modeling is then better.
- Recognition errors in each band should be lower, as there are less confusion possible between classes which depend only on the information present in each band: That implies that inputs of the recombination module contain less errors.
- As there are less inputs of the recombination module, the search space of the global optimum is less complex and training of the recombination module should then be better.

The first idea to build such classes is to study the confusion between phonemes in each band and to group the phonetic classes that are the most often confused. Such a work has been done by Mirghafori [Mirghafori99], in which she has studied three methods of grouping the classes: The first one makes use of the confusion matrices of the sub-recognizers, the second one makes use of the mutual information between classes and the third one attempts to minimize the error rate by choosing all possible groups of classes. Unfortunately, no major conclusions of this work has seemed to emerge. A possible reason is that it is too difficult to build these classes *a-priori.*

The other possibility to build these classes is to let the system adjust itself its sub-band classes. In the global training scheme, the classes in each band are built in order to optimize the final accuracy, and not the accuracy in each band, as it is the case in the two-steps training scheme. Using such a training procedure might lead to build classes in each band which are dependent of the information really presents in the band.

Finally, global optimization of a system, when it is possible, is always better than individual optimization of its components as the final accuracy depends not only on the intermediate components, but also on the relation which exists between them.

2. CHOICE OF THE OPTIMIZATION CRITERION

We are using the Minimum Classification Error (MCE) criterion to compute the parameters of the HMMs and the recombination model. We have chosen this criterion because it represents exactly our aim, that is to reduce classification error. Moreover, it has the advantage to be discriminant. We could not used the MLE criterion (Maximum Likelihood Estimation), as the outputs of the recombination module are not likelihoods. We could of course have derived a similar optimization scheme, aiming at « Maximizing the Score per Model », but we would then have lost the discrimination between models, and we have thought discrimination is very important in MBSR (intuitively, so that the models « specialize »

themselves to recognize some broad phonetic class during global training). Finally, this criterion has the advantage to be simple to implement in our system, which is not the case for the MAP (Maximum A Posteriori) and MMIE (Maximum Mutual Information Estimation) criteria. These two last criteria are difficult to implement for a single HMM, and they become almost impossible to manage when adding a recombination module.

3. PRINCIPLE

3.1. The MCE algorithm

When using the MCE criterion to train a system, one must first approximate the classification error function by a differentiable loss function, usually a sigmoïd:

$$l_{C(x)}(x) = \frac{1}{1+\exp(-\gamma\, d_{C(x)}(x))}$$

where γ is a constant, $C(x)$ is the real class of x, and $d_{C(x)}$ is a function quantifying the misclassification of a token. As defined in [Juang97], this function is typically:

$$d_{C(x)}(x) = -g_{C(x)}(x) + \left(\frac{1}{N-1} \sum_{j \neq C(x)} g_j^{\eta}(x) \right)^{1/\eta}$$

where N is the number of classes, η is a constant and $g_j(x)$ is the score returned by the system for class j and token x. When $\eta \to \infty$, this equation is simplified into:

$$d_{C(x)}(x) = -g_{C(x)}(x) + g_{\overline{C(x)}}(x)$$

$\overline{C(x)}$ represents the best class different from $C(x)$. We have used this approximation in our system.

The algorithm, described in details in [Juang97], applies then a gradient-descent procedure to the loss function. In each iteration, the parameters λ of the system are modifed by the following equation:

$$\lambda(t+1) = \lambda(t) - \epsilon \frac{\partial l_{C(x)}(x)}{\partial \lambda} \qquad \text{(Eq-1)}$$

3.2. Linear recombination

Let us now assume that the recombination module computes a score for each phone which is the weighted sum of the likelihoods returned by the corresponding models of all bands:

$$g_M(x) = \sum_{b=1}^{B} \alpha_{b,M} P(x|b,M)$$

To apply the MCE algorithm only to the recombination module, one simply makes use of:

$$\frac{\partial g_M(x)}{\partial \alpha_{b,M}} = P(x|b,M)$$

into Eq-1.

3.3. Non-linear recombination

Previous works have shown that best results were achieved in clean speech when a MLP is used in the recombination module. We have then adapted our global training model to implement such a recombination. Actually, few modifications are needed: The MCE criterion, when applied to a MLP, naturally transposes to the classical back-propagation algorithm, which is actually very similar to the MCE algorithm describes in section 3.1.

3.4. Modification of the HMM

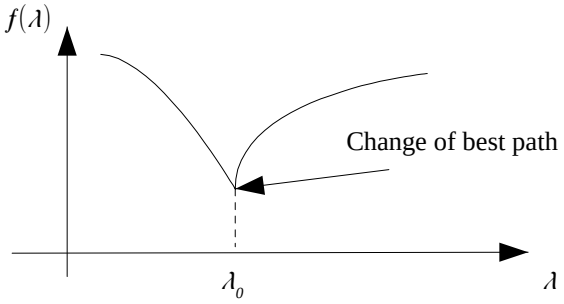
If we want to apply the same gradient-descent procedure to the sub-recognizers themselves, one has:

$$\frac{\partial g_M(x)}{\partial \lambda} = \alpha_{b,M} \frac{\partial P(x|b,M)}{\partial \lambda}$$

where λ is a parameter of the sub-recognizer.

$P(x|b,M)$ is not differentiable, as in our system, these likelihoods are computed by Hidden Markov Models. Actually, in HMMs, this likelihood depends on the best alignment frame-state which has been choosen. So, as long as λ stays in the range corresponding to the same best alignment,

$f(\lambda) = P(x|b,M$ is differentiable, but it is no longer the case when λ makes the best alignment change. Thus, $f(\lambda)$ is likely to have the following shape:



This can cause a problem if one wants, for instance, to decrease the score returned by a model: In this case, if λ is very close to λ_0, say $\lambda_0 - \delta$, the algorithm, wanting to decrease $f(\lambda)$, will move to $\lambda_0 + \delta$ which actually may increase $f(\lambda)$. However, we are not in the need of a differentiable function in order to increase or decrease $f(\lambda)$. The HMMs are trained by the Baum-Welch algorithm, which makes use of the counts of the number of times a Gaussian mixture or a transition is used. Hence, we just have to increase or decrease these counts in order to increase or decrease $f(\lambda)$. This algorithm is directly derived from the corrective training algorithm proposed by Bahl & al. [Bahl88] to enhance the training of the HMMs.

3.5. Overall Algorithm

The algorithm used to modify the parameters of the HMM is the following:

1. Initial training of all the HMMs is performed. The counts are saved in Γ(i) for model i.
2. The recombination module is initialized.
3. For each example u of the training corpus,
    4. The scores returned by each model for the example u are computed using the Viterbi algorithm. Let S(u,i) be the score returned by the model i for u. Let g be the good model of u.
    5. Compute the classification error E for u, using the scores S(u,i).
    6. Adjust the parameters of the recombination module using the error E, i.e. :
        7. If the recombination is linear, apply the gradient-descent procedure to the coefficients;
        8. If the recombination is neuronal, apply the back-propagation algorithm to the MLP.
    9. For each model w so that S(u,w) > S(u,g) - δ,
        10. Apply one iteration of the forward-backward algorithm to the model w on u. Let Γ(u,w) be the corresponding counts.
        11. Modify Γ(w) by : Γ(w) = Γ(w) - c Γ(u,w).
    12. Apply one iteration of the forward-backward algorithm to the model g on u. Let Γ(u,g) be the corresponding counts.
    13. Modify Γ(g) by : Γ(g) = Γ(g) + c' Γ(u,g).
    14. Adjust the HMM using the new counts Γ(i).
    15. Iterate the algorithm from step 3.

The coefficients c and c' which are used in step 11 and 13 adjust the importance of the modification of $f(\lambda)$. In our global training scheme, they are thus clearly related to the gradient of the classification error at the outputs of the HMM. So, the gradient descent procedure which is usually applied in the back-propagation algorithm must be applied in our algorithm "one step further", i.e. back to the outputs of the HMM. Then, the coefficients c and c' are set equal to this gradient multiplied by an empirically derived constant.

3.6. Initialization of the algorithm

The initialization of such a system can be as simple as the following:

1. Set all the HMM to the same « null » model;
2. In the case of linear recombination, set all the recombination coefficients to one;
3. In the case of neuronal recombination, set the initial weights of the MLP randomly.

This initialization is interesting, in the sense that no information which is dependent to only one of the two modules of our system is used. However, global training of a system is much longer than separate training, and the system might never converge. So, initial parameters are needed to be close of the global optimum. We have chosen to initialize the parameters as follows:

1. Perform classical training of the HMMs.
2. In the case of linear recombination, the coefficients are empirically chosen. In fact, they are set to 0.1 for the subbands and to 0.6 for the fullband;
3. In the case of neuronal recombination, an initial separate training of the MLP is performed.

4. EXPERIMENTAL RESULTS

4.1. Experimental setup

The Multi-Band system we are using is composed of five bands whose limits are [0 ... 538 Hz], [461 ... 1000 Hz], [923 ... 2823 Hz], [2374 ... 7983 Hz] and [0 ... 7983 Hz] (full-band). The fifth band is also the reference system and acoustic vectors are coded using 35 MFCC in this band. Actually, the signal is first passed to a filterbank which is composed of 24 triangular filters whose limits respect the mel scale. Then, 6 of these filters are assigned to each of the four sub-bands. This method computes so MFCC and not only cepstral coefficients as the original filterbank respects the Mel scale. Then, 3 cepstral coefficients are computed in each sub-band and 12 cepstral coefficients are computed in the full-band. Δ and Δ Δ coefficients are then added and the first coefficient is removed, which finally leads to 8 MFCC in each sub-band and 35 MFCC in the full-band.

The HMMs are second-order HMMs [Mari97] composed of three states, and a mixture of Gaussians is used in each state. The number of Gaussians is different in each state and is computed by the algorithm detailed in [Mari97].

4.2. Results

All these experiments are achieved in clean speech on the coretest part of the TIMIT database. *Figure 1* presents the results of two Multi-Band systems when global training is used. As the training part of the TIMIT database is very large (132500 examples), several days of computation are needed for each iteration, and the total number of iterations is then relatively low. Tests are done in isolated mode, i.e. manual segmentation of the signal is given to the system, which task is simply to associate to each segment a phone.

The reference system, the Multi-Band one using a linear recombination, and the Multi-Band one using a MLP have been simultaneously trained. In that way, the same amount of training has been used for the Multi-Band system as well as for the reference one. Moreover, training of all these systems has been achieved using the same training procedure derived from the MCE criterion which is described in section 3.5.

As these results demonstrate it, global training of the Multi-Band system leads to an improvement of the accuracy of this system. This may be due to several reasons:

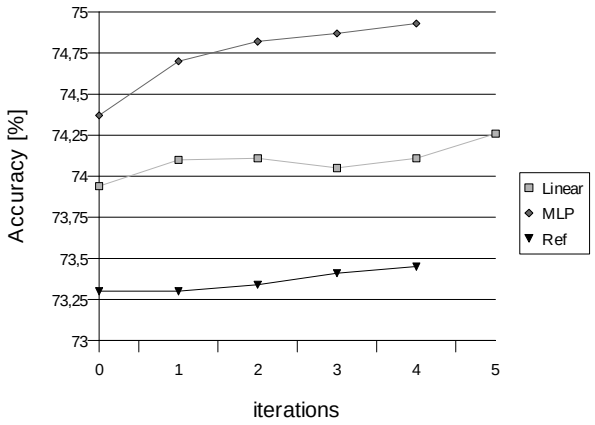


Figure 1: Accuracy of the globally trained Multi-Band systems in function of the number of iterations achieved during global training.

1. The global training algorithm tends to minimize the global classification error and modifies all the parameters of the system to reach its goal. That is not the case when the two-steps training algorithm is used, as this algorithm adjusts only the parameters of the recombination module to improve the final accuracy.
2. The acoustic models in the sub-bands and in the full-band are modified in order to minimize the final error rate. They might so represent the acoustic information really carried by the sub-bands in a better way than the classical phone models. These phone models are usually imposed in the sub-bands, despite the fact that we all agree to say that they are not adapted to sub-bands.

4.2. Modification of the classes

The modifications of the classes in the bands are quite complex to analyze, as many factors must be considered at the same time. To understand what happens in the phonetic space, we have studied the modification of the confusion matrix of each sub-recognizer during global training. We consider here a typical example: In the lowest band, one can see that class /ax/ (**a**bout) begins to absorb phone /ix/ (deb**i**t), i.e. more examples which are manually labelled as /ix/ are associated to the class /ax/ in the lowest band after global training than before. That can be understood as meaning that /ix/ is probably not essentially contained in the lowest band. Similarly, /ax/ is shrinking and disappearing from the fullband, which begins to specialize itself in the recognition of bursts. However, this is only a short example of what is happening in the phonetic space, and careful studies should be done to analyze the new roles of the classes. Actually, it might even be meaningless to try to analyze them in terms of classical phones, as they might be modeling other kinds of acoustic features and cues.

5. Conclusions and Future Works

We have proposed in this paper a new training algorithm for Multi-Band Speech Recognition Systems which globally optimizes the parameters of the system. This algorithm has significantly increased the accuracy and seems to be able to increase it much more, as all the parameters are not yet precisely tuned. Moreover, it is also very interesting as it modifies the classes that are recognized by the Hidden Markov Models in each band. This idea to use new classes in the frequency bands is very intuitive, but has not been successfully implemented until now. Global training might be one way to test it, as it does not impose a-priori classes, but it rather lets the system modify itself.

The immediate extension of this work is to study more precisely the new classes that have been generated. It should be very interesting to see if these modifications are similar to what one might intuitively predict, for example the grouping of vowels in the high frequency bands. This kind of work may then help us to design an architecture for our system which is more adapted to each band. Finaly, the study we have presented here about the global training algorithm is not yet finished, and several improvements may still be implemented, such as the use of other kinds of neural networks, or the possibility to change the number of classes in one sub-band, which is not yet allowed.

7. References

[Bahl88] **Bahl L. R., Brown P. F., de Souza P. V. and Mercer R. L.** : A New Algorithm for the Estimation of Hidden Markov Model Parameters. In *Proc. ICASSP'88*, April 1988.

[Bourlard97] **Bourlard H. and Dupont S.**: Subband-based speeh recognition. In Proc. *ICASSP'97*, Münich, Germany, pp. 1251-1254, 1997.

[Cerisara98] **Cerisara C., Haton J.-P., Mari J.-F. and Fohr D.**: A recombination model for multi-band speech recognition. ICASSP'98, Seattle, USA, mai 1998.

[Juang97] **Juang B.-H., Chou W. and Lee C.-H.**: Minimum Classification Error Rate Methods for Speech Recognition. In *IEEE Trans. on Speech and Audio Processing*, Vol. 5, N° 3, pp. 257-265, 1997.

[Mari97] **Mari J.-F., Haton J.-P. and Kriouile A.**: Automatic word recognition based on second-order hidden Markov Models. In *IEEE Trans. on Speech and Audio Processing*, Vol. 5, pp. 22-25, January 1997.

[Mirghafori99] **Mirghafori N. M.** : A Multi-Band Approach to Automatic Speech Recognition. Ph.D. thesis, International Computer Science Institute, Berkely, USA, janvier 1999.

[Tibrewala97] **Tibrewala S. and Hermansky H.** : Sub-band based recognition of noisy speech. In *Proc. ICASSP'97*, pp. 1255-1258, Münich, Germany, 1997.