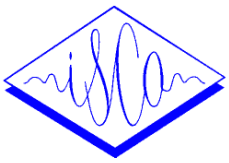

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

CODEBOOK WEIGHTS ADAPTATION FOR DISCRIMINATIVE TRAINING OF SCHMM-BASED SPEECH RECOGNITION SYSTEMS

C. Martín del Álamo (*), F. J. Caminero-Gil (*), C. de la Torre-Munilla, L. Hernández-Gómez (*)
cesar@gaps.ssr.upm.es, javi@gaps.ssr.upm.es, celinda@craso.tid.es, luish@gaps.ssr.upm.es

Speech Technology Group
Telefónica Investigación y Desarrollo
c/ Emilio Vargas 6
28043 - Madrid, SPAIN

ABSTRACT

Over the past few years, there have been several published reports on the use of Maximum Mutual Information (MMIE) for training HMM parameters. Lately, some reports have appeared, proposing different solutions to avoid the computational cost associated with the low convergence of the optimization technique [2] [5]. This paper proposes a new method for increasing the velocity of convergence, by pre-adapting the Codebook Weights of multiple codebooks in Semi-Continuous HMM (SCHMM). Experimental results on a small vocabulary recognition task show not only a fast convergence but also a 21% error rate reduction.

10.21437/Eurospeech.1995-23

1. INTRODUCTION

Many HMM-based speech recognition systems use multiple information streams in the computation of the output probabilities.

The use of multiple codebooks in SCHMMs (typically separating spectral or cepstral coefficients, their derivatives and some parameters related to the energy) allows individual quantification for each set of parameters. Multiple Codebooks improve the recognition accuracy because of a better representation of the particular characteristics of each stream.

Some authors [1] [2] [3] have obtained good results by training state-dependent stream weights in a discriminative way, that is, making use of a Minimum Error Rate criterion. However, from our point of view, there are still two related points that can be considered so as to improve the performance of previously proposed discriminative training procedures.

These two complementary points are:

1) To consider that, according to several studies [4][2], some parameters related to a particular stream can be more important than others represented in another stream. For example, in some states, modelling stationary zones, cepstral parameters seem to have a greater influence than their derivatives or than power, when discriminating the correct model, but this situation can be different in other zones.

2) In some cases, different behaviours in the different streams of a particular model can reduce the discrimination between models. For example, if the codebook of one stream does not fit well the data it will provide low probabilities an thus it will predominate over the rest of codebooks in the final probability. Thus, the discriminative possibilities between models will be greatly reduced.

In this paper, to overcome these drawbacks we present a two-step training procedure for SCHMM's.

1- The first step is called Codebook Weights Adaptation (CWA). We explicitly train, for each state of every model, the stream weights, trying to avoid the predominance of a poorly modelled stream. This training procedure makes use of a criterion based on the quantity of information provided by each stream.

2- For the second step, we use the algorithm proposed by Chen and Soong in [5], adapted to discriminative training of the state-dependent stream weights in SCHMM's.

The major results of the proposed procedure are:

a) A better performance of the two-step procedure compared to the single the discriminative training without a previous Codebook Weights Adaptation.

b) A faster convergence of the proposed two-step procedure compared to convergence of the discriminative training without a previous CWA.

The paper is organized as follows: Section 2 explains the Codebook Weights Adaptation procedure. The modification of the discriminative training proposed by Chen and Soong will be discussed in Section 3 and some results for a small-vocabulary isolated-speech recognition task are given in Section 4. A summary of conclusions are presented in Section 5.

2. CODEBOOK WEIGHTS ADAPTATION

When using multiple codebooks in HMM-based recognition systems, the output probability of a state j, given the observation O_t is obtained as:

$$P_j(O_t) = \prod_{s=1}^{S} [b_{js}(O_{ts})]^{w_{js}} \qquad (1)$$

Where b_{js} is the output probability of the stream s in state j. Assuming that the output distribution can be approximated as a linear combination of a set of gaussian distributions, shared by all the states, b_{js} can be represented by:

$$b_{js}(O_{ts}) = \sum_{m=1}^{M_s} c_{jsm} \cdot N(O_{ts}/\mu_{sm}, \Sigma_{sm}) \quad (2)$$

Where M_s is the number of mixtures per stream, weighted by the set of constants c_{jsm}, dependent of the state and stream. μ_{sm} and Σ_{sm} are the m^{th}-mixture mean and variance vectors, respectively.

The most common HMM parameter estimation technique is Maximum Likelihood Estimation (MLE) [10]. The objective function to maximize in MLE is:

$$R(\Theta) = \prod_{r=1}^{R} P_{\Theta}(\underline{O}^r/m_r) \quad (3)$$

Where Θ is the HMM parameter set, m_r is the model corresponding to the observation vector sequence $\underline{O}^r$ and R is the set of independent training observation sequences. $P_{\Theta}(\underline{O}^r/m_r)$ is the maximum *a posteriori* probability.

Using the forward-backward algorithm and the log likelihood formulation, the log of the objective function $LR(\Theta)$ for SCHMM's can be defined as:

$$LR(\Theta) = \sum_{r=1}^{R}\sum_{t=1}^{T_r}\left(\log U_j^r(t) + \sum_{s=1}^{S} w_{js} \cdot \log b_{js}\left(O^r{}_{ts}\right)\right) \quad (4)$$

Where T_r is the number of frames in utterance r, $U_j^r(t)$ stands for the probability of being at state j at time t, excluding the observation probability, for utterance r, that is:

$$U_j^r(t) = \left[\sum_{i=1}^{N} \alpha_i^r(t-1) \cdot a_{ij}\right] \cdot \beta_j^r(t) \quad (5)$$

and $\alpha_i(t)$ and $\beta_i(t)$ are the probabilities of the forward-backward algorithm. The summation over r covers all the utterances of the model in the training database.

To our knowledge, there is no criterion to include the codebook weights $\{w_{js}\}$ in the MLE of the HMM parameter set Θ. This is mainly due to the fact that including the restriction

$$\sum_{s=1}^{S} w_{js} = S \quad (6)$$

Maximizing $LR(\Theta)$ as function of codebook weights would lead to $w_{js} = S$ for the stream which provides the higher value of

$$\sum_{r=1}^{R}\sum_{t=1}^{T_r} \log b_{js}\left(O^r{}_{ts}\right) \quad (7)$$

and $w_{js} = 0$ for the other streams.

This result, although provides an obvious maximum in the objective function, do not seems reasonable.

In fact, as is usually stated when comparing discriminative training with MLE [2], it is not clear how an increase in (4) is related to the final objective of the system of reducing the error rate.

With this background we defined an empirical criterion to adapt the codebook weights. Our criterion is based on the assumption that the codebook weight of one stream should be proportional to its contribution into the objective function (4). That is, starting from w_{js}, the reestimated weights $\bar{w}_{js}$ will be:

$$\bar{w}_{js}^{-1} \propto \frac{\sum_{r=1}^{R}\sum_{t=1}^{T}\left(\log U_j^r(t) + w_{js} \cdot \log b_{js}\left(O^r{}_{ts}\right)\right)}{\sum_{r=1}^{R}\sum_{t=1}^{T}\left(\log U_j^r(t) + \sum_{n=1}^{S} w_{js} \cdot \log b_{js}\left(O^r{}_{ts}\right)\right)} \quad (8)$$

Where α denotes proportionality. The numerator in (8) can be considered related to the quantity of information provided by an specific stream while the denominator includes the contribution of all the streams. Note that the information from the transition probabilities is included in all the cases.

To adjust the proportionality factor we considered two possibilities:

1.- To derive its value from the restriction (6).

2.- To obtain a value for each stream as a function of the relationship between the contribution of the stream itself and the contribution of the transition probabilities.

Although this second possibility could be directed to reduce the mismatch between transition and output probabilities [2], we prefer for this preliminary work to leave this adaptation to the second, discriminative, training step. Then we used the first possibility.

Finally, as stated before, just because the proposed procedure do not guarantee an increase in the MLE objective function, we adapted the codebook weights after a standard MLE training. This previous MLE training was used to train means, variances and mixture weights of the SCHMM set. Afterwards, Codebook Weights Adaptation was done with very few iterations, where only codebook weights according to expression (8) were updated.

3. DISCRIMINATIVE TRAINING

As a second training step after standard MLE a Codebook Weights Adaptation described in the previous section, we included a discriminative training step which provides a more intensive relationship with the error rate. In the present work, we have used the algorithm proposed by Chen and Soong [5], extended for SCHMM's to be able to handle multiple codebooks.

In the following lines we will briefly describe the major extensions we made.

Firstly, a cost function is defined, based on the *frame-log difference*,

$$d_n(t) = \log b_j^n(O_t) - \log b_j^c(O_t) \tag{9}$$

where the superscript c means the correct state hypotesis, and the n means the n-best hypothesis.

The corresponding *frame-loss* function is defined as:

$$l_n(t) = \begin{cases} d_n(t) & \text{if } d_n(t) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

and the String-Loss difference:

$$L_n = \sum_{t=1}^{T} l_n(t) \tag{11}$$

which is the function to minimize.

We must remember that some constraints must be satisfied, such as the positive definiteness of the variance ($\sigma_{smi}^2 > 0$), the stochastic constraint for the mixture weights of each stream:

$$\sum_{m=1}^{M_s} c_{jsm} = 1 \tag{12}$$

and the restriction imposed in Eq. (6) for codebook weights.

As proposed in [5], for easier implementation, we used the following parameter transformation:

$$\begin{aligned} \bar{c}_{jsm} &= \log c_{jsm} \\ \bar{\sigma}_{smi} &= \log \sigma_{smi}^2 \\ \bar{w}_{js} &= \log w_{js} \end{aligned} \tag{13}$$

The derivatives of $d_n(t)$ to SCHMM parameters are:

$$\frac{\partial}{\partial \mu_{smi}} d_n(t) = w_{js} \cdot \gamma_{jsm} \cdot (O_{tsi} - \mu_{smi}) \cdot \sigma_{smi}^{-2} \tag{14}$$

$$\frac{\partial}{\partial \bar{\sigma}_{smi}} d_n(t) = w_{js} \cdot \gamma_{jsm} \cdot \frac{1}{2}\left[\frac{(O_{tsi} - \mu_{smi})^2}{\sigma_{smi}^2} - 1 \right] \tag{15}$$

$$\frac{\partial}{\partial \bar{c}_{jsm}} d_n(t) = w_{js} \cdot \gamma_{jsm} \tag{16}$$

$$\frac{\partial}{\partial \bar{w}_{js}} d_n(t) = w_{js} \cdot \log b_{jsm} \tag{17}$$

In the previous formulae,

$$\gamma_{jsm} = c_{jsm} \cdot N(O_{ts} / (\mu_{jsm}, \Sigma_{jsm})) \tag{18}$$

As can be seen, the expressions presented in [5] have been modified to introduce multiple codebook weights and shared mixture components.

After every iteration, inverse transformation is applied upon the transformed weights and variances, so that the aforementioned constraints are satisfied:

$$\begin{aligned} c_{jsm} &= \frac{e^{\bar{c}_{jsm}}}{\sum_{m=1}^{M_s} e^{\bar{c}_{jsm}}} \\ w_{js} &= \frac{e^{\bar{w}_{js}}}{\sum_{m=1}^{M_s} e^{\bar{w}_{js}}} \end{aligned} \tag{19}$$

This method was thought to work with N-best hypotheses, but in our work we used it only with one recognition result.

4. EXPERIMENTS

We have evaluated the proposed two-step training algorithm in a speaker-independent small vocabulary isolated speech recognition task through the telephone line.

The data we used is the VESTEL Speech Telephone Database (see [8] for global description). From VESTEL we selected only the isolated digits and command corpus and used a small vocabulary of 12 words: 10 digits plus the words "sí" (yes) and "no" (no).

The training set has 10150 utterances (700 per word), while the test set includes a separate set of 5464 utterances.

Firstly, a traditional MLE (Baum-Welch) training was applied, respecting the borders supplied by an endpoint detector [9]. In this stage, 4 mixtures-per-state continuous models were trained. From these mixtures, split in three different sections, 3 codebooks were constructed.

The first codebook (180 mixtures) includes 8 Mel Frequency Cepstral Coefficients (MFCC). The second

codebook (180 mixtures), includes the derivatives of the first codebook components (size = 8). And the third codebook (100 mixtures) combines the 0^{th} cepstral coefficient and its first derivative.

Once the SCHMM's were obtained, 5 reestimations of the mixture weights were made, reaching what we will call the Baseline System, which achieves an error rate of 2.53%.

From this point, in order to compare the proposed procedure with the original method, on one side, we applied the discriminative training, mentioned in section 3, to the Baseline System, updating the mixture weights, the means and the variances of the mixtures, and on the other side, we applied the Codebook Weights Adaptation, described in section 2, to the Baseline System, before the discriminative training.

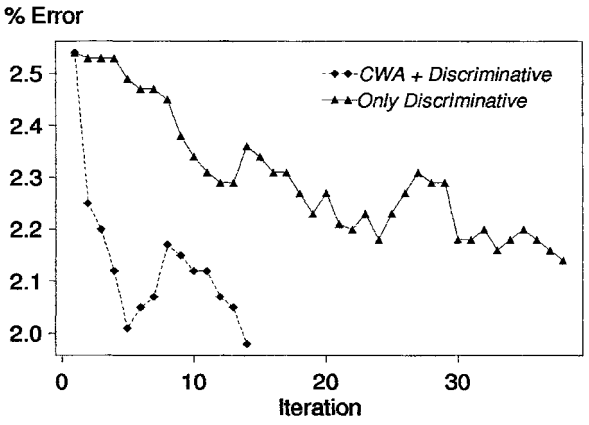

Figure. 1.- Error rate evolution on the test set.

By applying the discriminative training directly to the Baseline System, after 40 iterations, there were no errors left to train with, obtaining an error rate of 2.14%, which represents a 15% reduction from the baseline error rate.

On the other side, by applying the CWA to the baseline system, the error rate was decreased in a 10% (Error rate = 2.25%) with only 1 iteration. Afterwards, 1 iteration of discriminative training was given, reestimating only the stream weights, obtaining a new reduction of 2.4%. Then, as shown in Figure 1, with a few more iterations of discriminative training, reestimating means, variances and mixture weights, a best error rate of 1.98% was obtained (see Figure 1).

Thus, compared to the baseline system, the two-step training procedure achieves a 21% error rate reduction.

Note also that after CWA the convergence of the discriminative step was faster than when no CWA was used.

5. CONCLUSIONS

In this paper a new two-step training procedure is proposed for SCHMM's. This algorithm starts with the adaptation of the codebook weights based on the relative importance of each stream in the MLE objective function. After Codebook Weights Adaptation, discriminative training is used, looking for the minimization of the error rate. The recognition results on a small vocabulary recognition task indicate a significant performance improvement over discriminative training without previous codebook weights adaptation. The error rate, compared to MLE was improved from 2.53% to 2.14%, or a 15% reduction of word error using discriminative training, and from 2.53% to 1.98%, or a 21% reduction of word error using CWA before discriminative training. CWA experiments also indicate that the two-step training significantly improves the convergence on the discriminative phase. Thus, one of the major drawbacks in discriminative training, its slow convergence, is also addressed.

Further experiments on different recognition tasks, sub-word models and continuous speech, are under development to extend the results presented in this paper.

6. REFERENCES

[1] I. Rogina, A. Waibel, "Learning State-Dependent Stream Weights for Multi-Codebook HMM Speech Recognition Systems", ICASSP 94, Vol. I, pp. 217-220.

[2] Y. Normandin, R. Cardin, R. De Mori, "High-Performance Connected Digit Recognition Using Maximum Mutual Information Estimation". IEEE Trans. on Speech and Audio Processing, Vol. 2, NO.2, April 1994.

[3] F. Wolfertstetter, G. Ruske, "Discriminative State-Weighting in Hidden Markov Models", ICSLP 94, Japan.

[4] E.L. Boccheri and J.G. Wilpon, "Discriminative feature selection for speech recognition", Computer Speech and Language, Vol. 7, pp. 229-246 (1993).

[5] J.K. Chen, F.K. Soong, "An N-Best Candidates-Based Discriminative Training for Speech Recognition Applications", IEEE Trans. on Speech and Audio Processing, Vol. 2, NO. 1, PART II, pp. 206-216, January 1994.

[6] Cardin, Normandin and Millien, "Inter-Word Coarticulation Modelling and MMIE Training for Improved Connected Digit Recognition", ICASSP 93, Vol. II, pp. 243-246.

[7] Chou, Juang and Lee, "Segmental GPD Training Of HMM Based Speech Recognizer", ICASSP 92, Vol. I, pp. 473-476.

[8] D. Tapias, A. Acero, J. Esteve, J. Torrecilla, "The VESTEL telephone speech database", ICSLP-94, Japan. Sept. 1994.

[9] A. Acero, C. Crespo, C. de la Torre and J.C. Torrecilla, "Robust HMM-Based Endpoint Detector", EUROSPEECH-93, pp. 1551-1554.

[10] L.E. Baum, "An inequality with applications to statistical estimation for probabilistic functions of Markov processes", Inequalities, pp. 1-8, 1972.