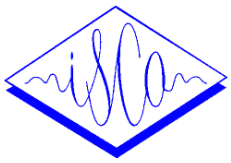

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

DISCRIMINATIVE TRAINING OF GARBAGE MODEL FOR NON-VOCABULARY UTTERANCE REJECTION

Celinda de la Torre and Alejandro Acero (*)

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

ABSTRACT

The aim of this paper is to describe a novel application of the *Discriminative Training* (DT) procedure for non-keyword rejection based on the principle of minimizing a *weighted error criterion*. This technique is applied to our Keyword Recognition System, a speaker-independent semicontinuous Density Hidden Markov Model (SCDHMM) recognizer. The proposed algorithm is evaluated for two different isolated word recognition tasks on telephone-line recordings containing both keywords and non-keywords utterances. We will compare the results with those obtained with *Maximum Likelihood Estimation* (MLE). In the Rejection Application the proposed procedure offers an automatic way of tunning the parameters for the desired application and a decrease on the Cost Function. Good results have been also obtained with the Discriminative Trained Garbage Model in the Word-Spotting Application.

1. INTRODUCTION

Interactive Voice Response (IVR) Systems are gaining popularity nowadays, because they allow users to interact with a remote computer by using simply a telephone, instead of requiring a more expensive and less portable computer equipped with a modem. Though many of IVR systems use the touch-tone as a means of user input, advances in speech recognition technology have made possible to use speech instead, particularly for those users which do not have touch-tone services. Although speech recognition technology also offers the capability of going beyond the recognition of the digits, and therefore opening a whole new realm of possibilities, the majority of the current IVR systems use speech recognition as a replacement for the touch-tone.

One of the problems encountered in field evaluation of IVR systems using speech recognition is that many users do not only utter the requested keywords, but also other words and noises (from hereon after non-keywords). When the user utters a non-keyword, it is up to the system to reject it, and when the user utters the keyword surrounded by other non-keywords, the system has to detect the keyword and ignore the rest (word-spotting).

Most techniques proposed in HMM-based speech recognition systems choose to use a so-called garbage model to cope with the rejection and word-spotting problems Typically, the garbage model is an HMM trained with all the non-keywords present in a database by traditional Maximum Likelihood Estimation (MLE) [2]. Since the garbage HMM model cannot accurately represent the acoustic variability of the non-keywords, its output log-probability has to be often empirically adjusted to obtain a better rejection rate, as shown in [1], where an affine transformation is used. Given the inability of the MLE techniques to generate good garbage models, some authors propose the use of *discriminative training* methods [4][5], which aim to minimization of recognition errors directly.

In this paper, we apply the *discriminative training* techniques only to the garbage HMM model, while maintaining the ML estimates for the keyword HMMs. We believe that discriminative training techniques will clearly outperform maximum likelihood techniques when the underlying acoustic model is not very appropriate, which is the case for the garbage HMM which is not designed to model many different acoustic events. We will evaluate this technique on a speaker-independent digit recognition task with long-distance telephone line recordings, and compare it with the results obtained with the ML estimated *garbage* model.

2. BASELINE RECOGNIZER AND EVALUATION METHOD

Our baseline system is designed to recognize speaker-independent isolated-speech (the ten spanish digits and a few control words) over long-distance telephone lines. Given that we want to be able to model noises before or after the keyword as well as to be able to have the system operate in a word-spotting fashion, our underlying system is a continuous-speech recognizer based on whole-word models and semicontinuous HMMs. The selected configuration consisted of 3 codebooks: 8 mel-cepstrums coefficients, their derivatives, and the power and its derivative with 180, 180 and 100 vectors respectively. We

(*) Currently working at Microsoft Corporation (One Microsoft Way, 9/1166. Redmond, WA 98052-6399)

used the Baum-Welch algorithm to obtain ML estimates for the keyword HMMs and the garbage HMM.

In order to evaluate the system we define a Cost Function (C) [1] that will be minimized. This Cost Function is a linear combination of the three main targets in an isolated Word-Recognizer with non-keyword rejection techniques: *Keyword Error Rate (Ek), Keyword Rejection Rate (Rk) and False Acceptance Rate (Fa).*

$$C = L_e \cdot E_k + L_r \cdot R_k + L_f \cdot F_a$$

with *weights*: $L_e + L_r + L_f = 1$

$$E_k = \frac{N_{ke}}{N_k}, R_k = \frac{N_{Re}}{N_k}, F_a = \frac{N_{Fa}}{N_g}$$

where:

N_k: Number of Keyword Utterances

N_g: Number of Garbages (non-Keyword Utterances)

N_{ke}: Number of Keyword Errors

N_{Re}: Number of Keyword Rejections

N_{Fa}: Number of False Alarm Rate:

3. DISCRIMINATIVE TRAINING OF THE GARBAGE MODEL

We can express the number of Keyword Rejections (N_{Re}) and the number of False Alarms (N_{Fa}) as:

$$N_{Re} = \sum_{N_k} U(P_g - P_k) = \sum_{N_k} U(\Delta_k)$$

$$N_{Fa} = \sum_{N_g} U(P_k - P_g) = \sum_{N_g} U(\Delta_g)$$

where:

Pg: Log-probability of the garbage HMM

Pk: Log-probability of the correct keyword HMM

U is the *Step Function.*

The discriminative procedure is used here to obtain the parameters of the garbage HMM that minimizes the *Cost Function (C).* Since the *Step Function* is not differentiable, we chose to approximate the N_{Re} and N_{Fa} expressions by a *Sigmoid Function.*

$$f(\Delta) = \frac{1}{(1 + exp(-T\Delta))}$$

$$\frac{\partial f}{\partial \Delta} = Tf(\Delta)(1 - f(\Delta))$$

where: $\Delta = \Delta_g$ in the F_a case

and $\Delta = \Delta_k$ in the R_k case.

The characteristics of this function and its derivative drive to take into account only those errors below a threshold (1/T), avoiding the influence of the large errors that probably could never be corrected and would produce undesirable effects. Moreover this approximation permits, not only to correct the errors between keyword and garbage models, but also to separated the scores between correct but easily confused pairs, the *near misses.*

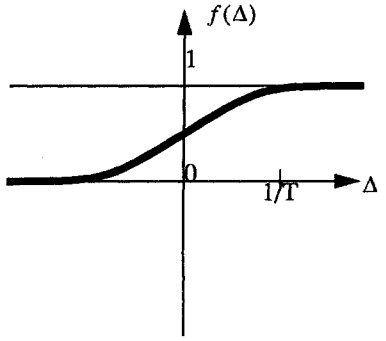

Figure 1: The Sigmoid Function

With this approximation, the cost function has the following form:

$$C = \frac{L_e}{N_k} \cdot N_{ke} + \frac{L_r}{N_k} \cdot \sum_{N_k} f(\Delta_k) + \frac{L_f}{N_g} \cdot \sum_{N_g} f(\Delta_g)$$

Since the effect of a different garbage model will be mainly seen in N*re* and N*fa*, when we take derivatives with respect to the garbage model parameters, we will assume that the contribution of Ek is zero, even though it is not mathematically true, because in practice we found that it does not make a difference and simplifies the calculations.

1. Implementation

Unlike other approaches [5], in our implementation only the parameters associated with the garbage model are discriminative trained. The problem of training the garbage model is that it should represent a broad class of acoustics events, so the Probability Density Functions of the garbage HMM trained with ML are not very sharp. For this reason the Garbage model has also high scores for the vocabulary words, resulting in an unbalanced low rejection and high False Acceptance Rate (See Table 1). A discriminatively trained garbage model obtained to minimize a Cost Function could balance these rates more effectively. Two set of parameters were trained:

- The *Weights of the Gaussian* ($p_{s,m}$) of the Probability Density Functions (pdf) (Codebook of the SCDHMM) where chosen in order to take advantage of the acoustic information provided by the codebook of the vocabulary.
- A set of *garbage-specific gaussians.* In order to represent some particular acoustic characteristics of the garbage, specially non-speech sounds, a specific set of gaussians were also discriminative trained and added to the vocabulary codebook. Those gaussians do not influence in the correct digit recognition.

For simplicity in the theoretical formulae and in the training procedure we have implemented the garbage by using only one pdf, formed by the weighted combination of vocabulary and garbage gaussians, where all the weights ($p_{s,m}$) and the complete garbage-specific gaussians (means

$\eta_{s,m,k}$ and variances $\sigma_{s,m,k}$) are discriminative trained.

This pdf can be easily implemented as a one state HMM, with the advantage of the low additional computational cost in the Recognition process.

2. Parameter Updating

The *Generalized Probabilistic Descendent Method* (GPD) [4] is used to obtain the parameter values (θ_i) that minimized the Cost Function:

$$\theta_{i_{DT}} = \theta_{i_{old}} - \beta \cdot \frac{\partial C}{\partial \theta_i}$$

$$\frac{\partial C}{\partial \theta_i} = \frac{\partial C}{\partial f} \cdot \frac{\partial f}{\partial \Delta} \cdot \frac{\partial \Delta}{\delta \theta_i}$$

The GPD Method has been theoretically justified to converge to a minimum of the Cost Function, although this minimum can be local. The convergence of the procedure is highly dependent on the *learning constant* (β) chosen. A big value of the *learning constant* can produce oscillation of the parameter values, which can means a divergence problem of the method. We chose *learning constants* to be proportionally to the decreasing magnitude of the *Cost Function*. A too small value of β could force the iterative procedure to need an excessive number of steps to converge to the minimum of C. Given our choice of initial values, we observed convergence after ten iterations.

Given a pronunciation:

$$\bar{O} = [O^1, O^2, ..., O^T]$$

where T is the number of frames of the pronunciation. We compute S different streams (one for static cepstrum, one for delta cepstrum and one for power and delta power):

$$O^t = [O_1^t, O_2^t, ..., O_S^t]$$

The accumulated log-probability generated by the one-state garbage model can be expressed as:

$$P_g(\bar{O}) = a_{1,1}^{T-1} \cdot \prod_{t=1}^{T} \prod_{s=1}^{S} \left[\sum_{m=1}^{M} p_{s,m} \cdot G_{s,m}(O_s^t) \right]^{w_s}$$

where M is the number of mixture gaussians of a stream and $p_{s,m}$ is the weight of the mixture m of the stream s ($G_{s,m}$).

The following are the final expressions of the derivatives of the accumulated log-probability of the garbage model respect to the three sets of parameters considered:

weight ($p_{S,M}$):

$$\frac{\partial}{\partial p_{S,M}} \ln P_g(\bar{O}) = w_S \sum_{t=1}^{T} \frac{G_{S,M}(O_S^t)}{\sum_{m=1}^{nmix} p_{S,m} G_{S,m}(O_S^t)}$$

mean ($\eta_{S,M,k}$):

$$\frac{\partial}{\partial \eta_{S,M,k}} \ln P_g(\bar{O}) = w_S \sum_{t=1}^{T} \frac{p_{S,M} \left(\frac{-1}{\sigma^2_{S,M,k}} (x^t_k - \eta_{S,M,k}) \right) G_{S,M}(O_S^t)}{\sum_{m=1}^{nmix} p_{S,m} G_{S,m}(O_S^t)}$$

variance ($\sigma_{S,M,k}$):

$$\frac{\partial}{\partial \sigma_{S,M,k}} \ln P_g(\bar{O}) = w_S \sum_{t=1}^{T} \frac{p_{S,M} \left(\frac{-1}{\sigma_{S,M,k}} + \left(\frac{1}{\sigma^3_{S,M,k}} (x^t_k - \eta_{S,M,k})^2 \right) \right) G_{S,M}(O_S^t)}{\sum_{m=1}^{nmix} p_{S,m} G_{S,m}(O_S^t)}$$

In practice, for each parameter set a different experimental *learning constant* has been used in order to consider independently the variation characteristic of each of them.

3. Smoothing

Since only incorrect recognized strings and near misses are taken into account in each iteration of the *Discriminative Training* process, the parameters obtained at each iteration should be smoothed with those from the previous iteration, trying to preserve the correctly recognized strings with the previous parameter set (θ_{old}). The Smoothing expression is:

$$\theta_{new} = (1-\alpha) \cdot \theta_{old} + \alpha \cdot \theta_{DT}$$

where $\alpha = \frac{R_k + F_a}{N_k + N_g}$, that is, the total misclassification rate.

This ensures the convergence avoiding large steps in the parameter values that could produce instabilities in the iterative procedure.

4. EXPERIMENTS AND RESULTS

To evaluate the training procedure we used an Isolated Word Recognition task with a vocabulary formed by the Spanish ten digits and four control words: "ayuda"(help), "repetir"(repeat), "comienzo"(begin) and "final"(end). The results we will describe for this task correspond to a situation when all fourteen words are simultaneously active. We also evaluated a Word-Spotting application.

1. Database Description

For the "ten digits + 4 control words" task, a large telephone database has been recorded throughout Spain, the Vestel Database [3]. For training (DV-T) a set of 1855C vocabulary utterances (digits + control words) and 584 out-of-vocabulary utterances and non-speech sounds was chosen. For testing (DV-R) the set was formed by 6358 vocabulary utterances and 456 out-of-vocabulary utterances and non-speech sounds.

To evaluate the Word-Spotting Application we used a Database (DV-W) formed by 2934 utterances containing keywords surrounded by non-keyword speech and noises. The training procedure has two different steps:

- In the first, a garbage model was training with a *Maximum Likelihood criterion* algorithm (MLE).
- In a second stage, the garbage model obtained in the first stage was iteratively estimated by the *Discriminative Procedure.*

The "Ten Digits + Control Words" task needed high rejection, so rejections are preferable than errors. For that reason the work point of the iterative *Discriminative*

Training procedure was obtained with the following values:

$$L_e = 0.6, L_r = 0.2, L_f = 0.2$$

Table 1 shows the results of the Rejection Application. The Cost Function has decreased moderately with the DT procedure, this means more than a 40% of improvement in the E_k, while obtained balanced R_k and F_a.

The results of the Word-Spotting Applications can be shown in Table 2. In this kind of applications the targets are two:

- The *Error Rate* (E_w) is the rate of keywords surrounded by out-of-vocabulary speech and noises that are incorrectly recognized as another keyword. It evaluates the ability of the garbage HMM in isolating the keyword for its recognition.
- The *Rejection Rate* (R_w) is the rate of vocabulary utterances with additional speech and noises that are recognized as *garbage* and rejected.

The Cost Function can easily be extended to the case of Word-Spotting Applications. Two new constants (L_{we} and L_{wr}) are chosen to weight the Word-Spotting targets in C. Their contribution should be proportional to the percentage of Word-Spotting utterances (P_w) in a real system. We took this percentage from a Field Trial done with our "News Service" system [6], which shown than a 10% of the responses were vocabulary words surrounded by out-of-vocabulary words and noises.

$$C = (1-P_w)\,(L_eE_k + L_rR_k + L_fF_a) + P_w\,(L_{we}E_W + L_{wr}R_W)$$

with $L_{we} = 0.8, L_{wr} = 0.2$ and $P_w = 0.1$.

The goodness of the *garbage model* trained with DT is proven with the results obtained in the Word-Spotting Application. This demonstrated that a HMM training with the proposed procedure gives good results in applications different than the one in which it was trained.

Method	C	E_k	R_k	F_a
MLE	5.22	1.8%	2%	18.6%
DT	4.76	1.0%	8.9%	11.8%

Table 1: Ten Digits + Control Words. Rejection Results. (DV-R)

Method	C	E_w	R_w
MLE	5.50	9.8%	0.8%
DT	5.07	6.9%	12.0%

Table 2: Ten Digits + Control Words. Word-Spotting Results. (DV-W)

The longer duration and the phonetic characteristics of the control words make of them a high confusable set with respect to the out-of-vocabulary one, that means that the Rejection and False Alarm Detection tasks are more difficult.

5. DISCUSSION AND FUTURE WORK

The discriminative method described (DT) is able to balance the Rejection Rate and False Alarm Rate so as to minimize the compound *Cost Function*, whereas the MLE method is not designed to balance them. Therefore it is not surprising that the DT method yields a lower cost function than the MLE method (See Table 1). We were disappointed, however, that the decrease in cost function was only 8% relatively. This points out the need for a more elaborated *garbage* model, with more states.

Our current work is focused in studying new *Decision Rules* for our *Discriminative Training Procedure* in order to make the implementation independent of the discriminative task and general for any HMM topology.

6. CONCLUSION

In this paper we have presented a novel application of the *Discriminative Training* (DT) procedure for non-keyword Rejection. The procedure is consistent with the recognition framework with a very low additional computational cost in the recognition task. The garbage parameters are automatically trained with the Discriminative Procedure. The more simple implementation of the method, a one-state garbage HMM, has provided low and balanced Keyword Rejection and False Alarm Rates in the Rejection Application. The garbage model obtained has also shown good behaviour in the Word-Spotting Application.

ACKNOWLEDGMENTS

We are very grateful to all the members of the Speech Technology Group of Telefonica I+D, specially to those working with us in the HMM Recognition System.

REFERENCES

[1] L. Villarrubia and A. Acero. "Rejection Techniques for Digits Recognition in Telecommunication Applications".Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing. Minneapolis, April 1993, pp 455-458.

[2] J. G. Wilpon, L. G. Miller and P. Modi. "Improvements and Application for Keyword Recognition using Hidden Markov Modelling Techniques". Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing.Toronto, Canada, May 1991. Pp 309-312

[3] D. Tapias, A. Acero, J. Esteve and J.C. Torrecilla. "The Vestel Telephone Speech Data Base". Proc. Int. Conf. on Spoken Language Processing, Yokohama, Sept. 94.

[4] W. Chou, B.H. Juang and C. H. Lee. "Segmental GPD Training of HMM based Speech recognizer". Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing. San Francisco, March 1992, pp 473-476.

[5] R.C. Rose. "Discriminant Word Spotting Techniques for Rejecting Non-Vocabulary Utterances in Unconstrained Speech". Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing. San Francisco, March 1992, pp 105-108.

[6] M.J. Poza, C. de la Torre, D. Tapias and L. Villarrubia. "An Approach to Automatic Recognition of Keywords in unconstrained Speech using parametric models". Proc. Eurospeech 91. Genoa, September 91,pp 471-474.