ACOUSTIC SPACE DIMENSIONALITY SELECTION AND COMBINATION USING THE MAXIMUM ENTROPY PRINCIPLE

Yasser H. Abdel-Haleem[*]*, Steve Renals*[†]*, Neil D. Lawrence*

Department of Computer Science,
Sheffield University, 211 Portobello Street,
Sheffield S1 4DP, UK.
{y.hifny, n.lawrence}@dcs.shef.ac.uk, s.renals@ed.ac.uk

ABSTRACT

In this paper we propose a discriminative approach to acoustic space dimensionality selection based on maximum entropy modelling. We form a set of constraints by composing the acoustic space with the space of phone classes, and use a continuous feature formulation of maximum entropy modelling to select an optimal feature set. The suggested approach has two steps: (1) the selection of the best acoustic space that efficiently and economically represents the acoustic data and its variability; (2) the combination of selected acoustic features in the maximum entropy framework to estimate the posterior probabilities over the phonetic labels given the acoustic input. Specific contributions of this paper include a parameter estimation algorithm (generalized improved iterative scaling) that enables the use of negative features, the parameterization of constraint functions using Gaussian mixture models, and experimental results using the TIMIT database.

1. INTRODUCTION

The maximum entropy (MaxEnt) principle encourages us to choose the most unbiased distribution that is simultaneously consistent with a set of constraints. Typically, the available information about the system is incomplete, and there is an infinite number of possible probability distributions that satisfy the constraints. E. T. Jaynes suggested maximizing Shannon's entropy criterion subject to the given constraints to choose a suitable distribution as follows [1]:

> When we make inferences based on incomplete information, we should draw them from that probability distribution that has the maximum entropy permitted by the information we do have.

Recently, MaxEnt has been used in the field of Natural Language Processing (NLP) as a principled way to combine multiple sources in a probabilistic framework [2]. In speech recognition, MaxEnt has been applied to language modelling [3], but there has been relatively little work in acoustic modelling: Likhododev and Gao [4] developed a direct model for speech recognition whose parameters were estimated by MaxEnt, and Macherey and Ney [5] discriminatively estimated the parameters of a Gaussian model based speech recognizer using MaxEnt.

In this paper, we propose a general methodology for discriminant dimensionality reduction. By discriminant analysis we mean that the training samples are labelled according their class membership. Hence, it is a supervised machine learning approach, which is developed for a continuous feature space. The acoustic space is represented by a large number of acoustic features that have been developed specifically for speech recognition. This large number of acoustic features is searched to find a set of features that are optimal in terms of phoneme class separation. The selected acoustic features are combined in the MaxEnt framework to estimate the posterior probabilities over the phonetic labels given the acoustic input. The MaxEnt constraints will be defined as the expectations of the transformed acoustic features collected from the acoustic space. An important consequence of this approach is that it is possible to infer optimal feature sets on a per-class basis, while ensuring the comparability of distributions between classes by summing over the full acoustic feature space.

In the next section, a mathematical treatment for the principle of Maximum Entropy is presented and in section 3 we introduce the parameter estimation procedure. Section 4 discusses how to define and implement the acoustic constraints. In addition, the process of dimensionality reduction and selection in the MaxEnt framework is reviewed in section 5. Section 6 introduces the experimental work and preliminary results on the TIMIT database. We conclude and discuss further work in section 7.

2. THE MAXIMUM ENTROPY PRINCIPLE

Let y be a discrete variable representing the possible output classes in a classification problem, and x be an observation affecting the states of the system. The constrained optimization problem in hand is to maximize the conditional Shannon entropy:

$$\arg\max_{p\in\mathcal{C}} S(p) = -\sum_x \tilde{p}(x)\sum_y p_\Lambda(y \mid x)\ln p_\Lambda(y \mid x) \qquad (1)$$

subject to

C1 $p_\Lambda(y \mid x) \geq 0$ for all y, x. and $\sum_y p_\Lambda(y \mid x) = 1$ for all x.

C2 $\sum_x p(x)\sum_y p_\Lambda(y \mid x) g_i(x,y) = \sum_{x,y} \tilde{p}(x,y) g_i(x,y) = \tilde{p}(g_i)$ for $i = 1, 2, \ldots, n$.

Where $S(p)$ is the expectation of the conditional entropy of the model with respect to the training database, $\tilde{p}(x)$ is the observed marginal probability, and $\Lambda = \{\lambda_i\}$ is the set of parameters to be optimized. Constraint C1 represents the direct constraint from probability theory. Constraint C2 represents the integration of the available prior knowledge on the random variables x, y in terms of

[*]Yasser H. Abdel-Haleem is sponsored by a Motorola Studentship.
[†]Now at CSTR, University of Edinburgh, Edinburgh EH8 9LW, UK

the characterizing constraints $g_i(x, y)$, which have expected value $\tilde{p}(g_i)$.

The maximum entropy problem formalism results in a probability distribution, which is the log linear or exponential model:

$$p_\Lambda(y \mid x) = \frac{1}{Z_\Lambda(x)} \exp\Big(\sum_i \lambda_i g_i(x, y)\Big) \quad (2)$$

Where

- λ_i is the Lagrange multiplier (weighting factor) associated to the function $g_i(x, y)$.
- $Z_\Lambda(x)$ (Zustandsumme) is a normalization coefficient resulting from the natural constraints over the probabilities summation, commonly called the partition function, and given by

$$Z_\Lambda(x) = \sum_y \exp\Big(\sum_i \lambda_i g_i(x, y)\Big)$$

The entropy is a concave function of the mean values of the characterizing constraints $\tilde{p}(g_i)$ [6]. Hence, the MaxEnt solution is unique given the empirical mean values of the constraints. Practically this means that the solution is not sensitive to the initial values of the model parameters and the constructed model is unique for a given database in the statistical learning procedure.

It should be noticed that in the absence of any constraint except the natural constraint, the maximum entropy formalism results in the flat uniform distribution:

$$p_\Lambda(y \mid x) = 1/n. \quad (3)$$

This result explains the basic philosophy behind maximizing the entropy as the uniform distribution is the most unbiased distribution. Integrating constraints results in reduction of the entropy but the output distribution is the most unbiased distribution consistent with constraints.

Consider a maximum entropy problem with two constraints μ and σ^2 of a continuous random variable whose probability density function is square-integrable. In such a case, when the continuous entropy is maximized, its solution is the normal distribution. This explains the importance of this distribution and why it has been frequently used in the application of statistical inference and why it deserves the adjective "normal", where this distribution is the most uncertain and maximizes the entropy [7]. The strong assumption that the data is normally distributed for the two constraints μ and σ^2 is relaxed by introducing the concept of the parametric constraints in section 4.

3. PARAMETER ESTIMATION

The purpose of the parameter estimation algorithm is to estimate the parameters $\lambda_1..\lambda_n$ using numerical methods. A modified version of the Improved Iterative Scaling (IIS) algorithm [8] was used to estimate the parameters. It was suggested to us by John Lafferty [9] to support constraints that may take negative values, which was a restriction of the original algorithm. Further details about the mathematical derivation are reported in [10]. The basic idea behind the IIS algorithm is to make use of an auxiliary function, which bounds the change in divergence from below after each iteration.

The Generalized Improved Iterative Scaling (GIIS) algorithm proceeds as follows:

1. Let $p^0_\Lambda(y \mid x)$=$1/n$, which is the uniform model, where $\lambda_i = 0.0$. $i = 1, 2, \ldots, n$
2. Solve the following equation using Newton's method:
$\tilde{p}(g_i) =$
$\sum_x \tilde{p}(x) \sum_y p^t_\Lambda(y \mid x) g_i(x, y) \exp\big(\delta_i^{t+1} s_i(x, y) M(x, y)\big)$
3. Update the parameters: $\lambda_i^{t+1} = \lambda_i^t + \delta_i^{t+1}$
4. If a valid termination condition is achieved then stop else go to step 2.

Where $M_i(x, y) = \sum_i |g_i(x, y)|$ and $s_i(x, y)$ is the sign of $g_i(x, y)$. Solving the equation in step 2 for each iteration, results in the value of Maximum Likelihood (ML) step δ_i^{t+1} towards obtaining the MaxEnt global solution. When $s_i(x, y)$ is positive, step 2 corresponds to the IIS algorithm. Furthermore, it can be shown that the equation has a unique solution by directly checking its convexity.

4. PARAMETRIC CONSTRAINTS

The description of the constraining characterizing functions is an optional implementation issue in which the prior knowledge for different applications is integrated. The characterizing functions are expected to have different values for different classes and observations. Hence, the estimated posterior probabilities will have a meaning over classes. Here we introduce the idea of parametric constraints to enable the flexible modelling for constraints based on continuous observations.

These parametric constraints aim to model the high variability of the observed acoustic features and overcome the strong assumption that the data distribution is Gaussian if we used the acoustic features directly. The form of the parametric constraints is optional: in this work we have used finite GMMs, which are a flexible model with a strong and rich history in speech recognition. The GMMs are estimated per acoustic feature per label using the EM algorithm [11]. The resulting conditional mixtures will estimate the soft log likelihood score for an acoustic feature, which will take the role of MaxEnt constraints over labels classes per event. The GMMs constraints have the following form:

$$g_i(x, y) = g_i(x, y; \theta) = P_\theta(x \mid y) \quad (4)$$

Where

- x is the observed continuous random variable. In acoustic space, it represents the acoustic features values per frame.
- y is a discrete random variable represents output classes.
- $P_\theta(x \mid y)$ is the likelihood score for conditional GMMs.

The Gaussian mixture is defined as a convex combination of Gaussian densities. A Gaussian density in a d-dimensional space, characterized by its mean $\mu \in R^d$ and d x d covariance matrix Σ is defined

$$\phi_\theta(x \mid y) = \frac{1}{\sqrt{2\pi}^d \sqrt{\det(\Sigma_y)}} e^{-\frac{1}{2}(x-\mu_y)^T \Sigma_y^{-1}(x-\mu_y)} \quad (5)$$

where θ denotes the parameters μ and Σ. A k component GMM is then defined as:

$$P_\theta(x \mid y) = \sum_{j=1}^{k} q_j \phi_{\theta j}(x \mid y) \quad (6)$$

with $\sum_{j=1}^{k} q_j = 1$ and for $j \in \{1, \ldots, k\} : q_j \geq 0$.

5. DISCRIMINATIVE CONSTRAINT SELECTION AND COMBINATION

In the MaxEnt solution, the Lagrange multipliers, which may be interpreted as the importance of each constraint, are the outcome of the training procedure. In many cases, the number of the available constraints may be large. Hence, the estimation of the MaxEnt model parameters may be computationally intensive or impractical. Obtaining the MaxEnt solution in incremental steps is a practical way to evaluate the importance of the evaluated constraints. In particular, this methodology may be considered as dimensionality reduction and selection. Unfortunately, evaluating the importance of every constraint by building a MaxEnt model incrementally will invalidate the previous estimate of the model parameters. Thus all the model parameters must be re-estimated at each step, which is computationally intensive.

Della Pietra et al [8] developed an efficient solution to the problem in which the Lagrange multipliers of the constraints are kept fixed while evaluating a given constraint. This yields a great computational saving as the problem is reduced to one dimensional optimization problem. This inductive approach is based on a measure referred to as the constraint gain, where the gain of a constraint is usually measured in terms of the increase of the log-likelihood of the training data by adding the evaluated constraint:

$$\begin{aligned} \text{Gain}_{g_i}(\beta) &= \triangle L_{g_i} \\ &\approx L(\hat{p}_{g_i}(y \mid x)) - L(p(y|x)) \\ &\approx \sum_x \tilde{p}(x) \log Z_{g_i}(x) + \beta \tilde{p}(g_i) \end{aligned} \tag{7}$$

where $\hat{p}_{g_i}(y \mid x)$ is an approximate MaxEnt model constructed after adding the constraint g_i and β is the estimated Lagrange multiplier for an evaluated constraint associated with the function $g_i(x, y)$.

6. EXPERIMENTAL WORK

We have performed experiments using the TIMIT database. In these experiments we have used the TIMIT phone labels as the classes in the MaxEnt model. The 61 phone classes in TIMIT were reduced to a set of 39 labels in the standard way. We used the 420 speaker training set, analyzed using a 32ms Hamming window at a 16 ms fixed frame rate, resulting in 880 564 frames.

We extracted a large number of acoustic features: MFCCs, PLP, and RASTA-PLP, along with the first and second order derivative features for each set. This resulted in a total of 117 (39×3) acoustic features. The set of MaxEnt constraints were then obtained by taking the product of the class and acoustic feature spaces, resulting in a total of 4563 (39×117) MaxEnt constraints:

$$g_{y'}(x, y) = \begin{cases} P_\theta(x \mid y) & \text{if } y' = y \\ 0 & \text{otherwise} \end{cases}$$

The parametric GMMs constraints were estimated for each acoustic feature (one dimensional GMMs). Each GMM had four Gaussian components. Incremental constraint selection was applied to select the best 1600 acoustic constraints. This is approximately equivalent to the number of constraints in a standard system with 39 phone classes and 39 acoustic features. In this work we aimed to find the optimal set of constraints, without requiring an equal number of acoustic features per phone.

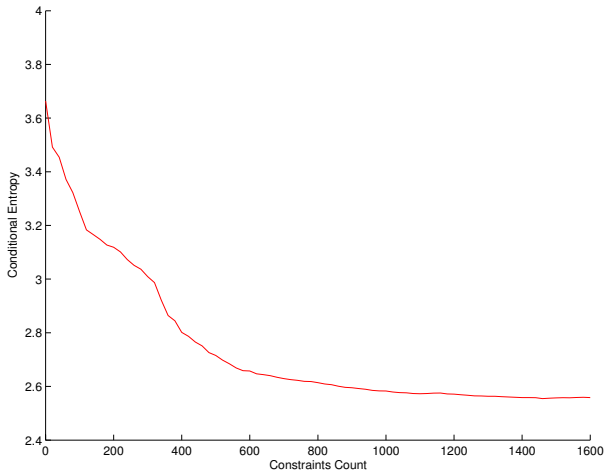

Fig. 1. The reduction of conditional entropy due to the incremental integration of the acoustic constraints.

The corpus was sampled to calculate the model expectation using the training procedure described in section 3. In addition, at each incremental step the best 20 constraints were selected and the model parameters re-estimated. As shown in figure 1, integrating new constraints results in reduction of the conditional entropy. This implies an increase in the likelihood of the model with respect to the training data. The reduction of the entropy became insignificant after adding about 1200 constraints, corresponding to an average 32 features per phone. The silence phone was not included in the constraint selection process.

Figure 2 shows the number of acoustic features (constraints) selected for each phone class. Since 1600 constraints were selected, the expected number of features per class was 40 (2.5%). Those phones with a high degree of variation were represented by more features, selected by the discriminative process. For instance, the unvoiced fricative /s/ was represented by 96 constraints (6% of the total).

Most of the three types of the acoustic features have been chosen during the selection process to represent different phone classes. The PLP features represent approximately about 40% of the selected constraints. The MFCC constraints represent about 33% of the selected features. The low coefficients (plp01, plp02, mfcc01, mfcc02 etc..) from the acoustic features are strongly selected during the selection process as shown in figure 3. The same behaviour is noticed for the low order coefficients of the derivatives of the acoustic features.

7. CONCLUSION

In this paper we present an approach for discriminative feature selection based on MaxEnt modelling, and have demonstrated an application to dimensionality reduction in the acoustic space. In this approach the acoustic observations per class were formulated as parametric constraints, using GMMs. This aims to relax the strong assumption that the data is normally distributed if the acoustic features are used directly. The MaxEnt principle addresses two fundamental questions: What are the important features required to model the acoustic information? How should these features be

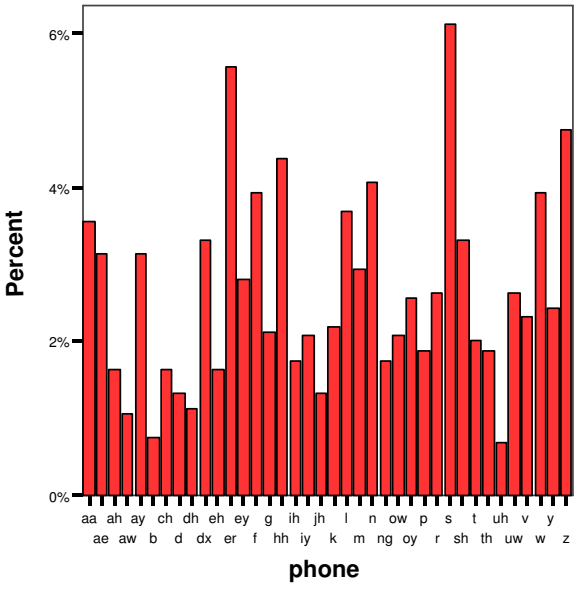

Fig. 2. Variable acoustic feature count for each phone class.

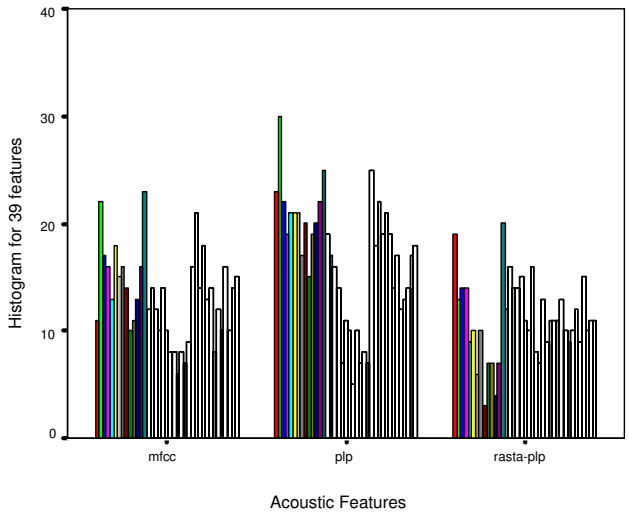

Fig. 3. Relative usage for different acoustic features during the selection process.

combined? We have shown that phonetic classes can be modelled with variable length acoustic feature vectors, selected automatically using the discriminative MaxEnt framework. Our future work will concentrate on optimal and efficient training and the integration of estimated posterior probabilities of MaxEnt Models within HMM based systems for continuous speech recognition.

8. REFERENCES

[1] E. T. Jaynes, "On the rationale of maximum-entropy methods," *Proc. IEEE*, vol. 70, no. 9, pp. 939–952, 1982.

[2] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[3] R. Rosenfeld, "A maximum entropy approach to adaptive statistical language modeling," *Computer, Speech and Language*, vol. 10, pp. 187– 228, 1996.

[4] A. Likhododev and Y. Gao, "Direct models for phoneme recognition," in *Proc. IEEE ICASSP*, 2002, pp. 89–92.

[5] W. Macherey and H. Ney, "A comparative study on maximum entropy and discriminative training for acoustic modeling in automatic speech recognition," in *Proc. Eurospeech*, 2003, pp. 493–496.

[6] J.N. Kapur and H.K. Kesavan, *Entropy Optimization Principles with Applications*, Academic Press, 1992.

[7] S. Guiasu and A. Shenitzer, "The principle of maximum entropy," *The Mathematical Intelligencer*, vol. 7, no. 1, 1985.

[8] Stephen Della Pietra, Vincent J. Della Pietra, and John D. Lafferty, "Inducing features of random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 4, pp. 380–393, 1997.

[9] J. Lafferty, "Personal communication," May 2002.

[10] Yasser H. Abdel-Haleem, "The use of maximum entropy principle in continuous speech recognition," May 2003, http://www.dcs.shef.ac.uk/~yhifny/publications/MaxEnt-ASR.pdf.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the *EM* algorithm," *Journal of the Royal Statistical Society B*, vol. 39, no. 1, pp. 1–38, 1977.