Maximum Likelihood Learning of Auditory Feature Maps for Stationary Vowels

Kuansan Wang *Chin-Hui Lee* *Biing-Hwang Juang*

Speech Recognition Department
Lucent Technologies Bell Laboratories
600 Mountain Avenue, Murray Hill, NJ 07974

ABSTRACT

In this paper, a mathematical framework for learning the acoustic features from a central auditory representation is presented. We adopt a statistical approach that models the learning process as to achieve a maximum likelihood estimation of the signal distribution. An algorithm, called *statistical matching pursuit (SMP)*, is introduced to identify regions on the cortical surface where the features for each sound class are most prominent. We model the features with distributions of Gaussian mixture densities, and employ the expectation-maximization (EM) procedure to both improve the parameterization and refine iteratively the selection of cortical regions from which the features are extracted. The learning algorithm is applied to vowel classification on TIMIT database where all the vowels (excluding diphthongs, nine in total) are regarded as individual classes. Experimental results show that models trained under SMP/EM algorithm achieve a comparable recognition accuracy to that of conventional recognizers.

1. Introduction

The problem of pattern recognition, according to Bayes [4], can be succinctly stated as the problem of implementing a *maximum a posteriori (MAP)* probability decision scheme. Bayes' approach requires that the *a posteriori* probability be known for any observed pattern, say x. The difficulty in practice lies in the fact that this probability or even its functional form is usually unknown. Over the past few years, mixture densities (particularly Gaussian mixtures) have found widespread use for modeling the class conditional density, $p(x|C)$. This is mostly because it provides a convenient combination of the advantages of a parametric and a non-parametric distribution. Class conditional densities are normally modeled separately from the class prior $p(C)$; combining the two easily leads to the needed *a posteriori* probability.

The use of mixture densities does not entirely solve the problem in choosing the right model for the data. There are several remaining issues. First, there is a tendency, the amount of training data permitting, to use models with a large number of parameters, approaching a non-parametric one in effect. In automatic speech recognition (ASR), this is a convenient way to improve the recognizers' accuracy performance, with a cost of having to collect more training data. However, increasing the model size eventually compromises the generalization capability of the model. Second, the raw observation or measurement of a pattern is usually expressed in a high dimensional space. Designing a recognizer in a high dimensional space in which the classes are not necessarily separable either leads to inefficient models (i.e., more parameters are required than necessary) or causes practical difficulties in evaluating the probabilities and implementing the decision process. This is particularly so when sequential search algorithms are involved, as in the case of ASR. There is thus a need to reduce the raw observations for large scale recognition tasks. Third, the pattern observation is often subject to contamination or distortion which will inevitably change the form of the data distribution. For a recognizer to function equally well under adverse conditions, the transformation process, from the observation to feature, must bring out the fundamental characteristics of the pattern that is ideally invariant or less susceptible to contamination or distortion. Finally, it is imperative that any approach for feature extraction must be geared toward optimizing the overall recognition performance. This dictates the design of a feature extraction algorithm be subject to the same set of evaluation criteria as the back end classifier, since, obviously, the contributions of these two subsystems are hardly separable in the overall recognition performance.

In terms of signal processing and feature extraction methods, auditory approaches have inspired many algorithms that render profound impacts on speech applications. Among the well known examples, mel-scale spectrum and the use of dynamic features (such as delta cepstrum) can all relate their motivations to certain auditory processing strategies. In this paper, we present our recent effort in adapting to the ASR applications the feature mapping mechanism in the central auditory system. Our goal here is to formulate the auditory mechanism into a viable feature extraction algorithm. The auditory approach, as will be briefly summarized later, consists of highly elaborated mechanisms in analyzing the acoustic power spectrum. The challenge of this effort, then, is to maintain a balance in preserving the elaborated nature of the auditory approach, while, in the mean time, keeping the algorithmic complexity at a manageable level.

The auditory processing studied in this work is based on the recent physiological findings in the primary cortex. It is shown [7] that, at the cortical level, the auditory system seems to perform a windowed Fourier analysis on the (short-time) power spectrum of the

incoming signal. From an algorithmic point of view, the processing bears remarkable similarity to what the peripheral auditory system imposes on the acoustic waveform. More specifically, the neural responses on the cortical surface follow closely the Fourier transforms of the acoustic power spectrum windowed at various frequencies. In a rather precise analogy, as a spectrogram is a two dimensional map depicting the windowed Fourier transform of an acoustic waveform, the cortical feature map is basically a *cepstrogram* of the power spectrum. While the auditory approach resembles in spirits to the conventional cepstrum analysis in computing the Fourier transform of the power spectrum, it bears a critical distinction: in the auditory approach, the spectral "locality" of acoustic features is preserved. As there is overwhelming evidence that salient speech features reside locally in the frequency domain [1], preserving frequency-local features may well prove to be a very important property for feature extraction.

To accommodate this representation into an ASR system, the following procedures are taken. First, we model the cortical response as a linear projection of the power spectrum onto a collection of *analyzing functions*, in which each analyzing function is a neuronal response function from the cortical model described in [7]. We introduce an iterative algorithm, called *stochastic matching pursuit (SMP)*, to extract features from the elaborated representation. SMP algorithm is designed to select a minimal subset of the analyzing functions whose responses can efficiently represent the statistical properties of the signals for each recognition class. The selection of a set of analyzing functions can be viewed as choosing regions on the cortical surface (a feature map) in which the neuronal responses correspond to the chosen analyzing functions. The selection criterion is included in the training procedure, in which a single optimality function for the classifier *and* feature extraction can be applied. Within the scope of this paper, we employ the maximum likelihood (ML) criterion commonly seen in many ASR training procedures. It can be shown that the proposed learning algorithm converges unconditionally. From a theoretical viewpoint on pattern recognition, the algorithm enforces a distinctive link between feature extraction and pattern parameterization such that the optimality criteria for these two stages are congruent with each other.

2. Stochastic Matching Pursuit Algorithm

In this section, we present the detailed formulation of SMP algorithm. SMP algorithm bases its fundamental framework on the matching pursuit algorithm proposed by Mallat *et al.* [5]. The extension augmented in SMP is modest: while the original matching pursuit algorithm is designed to decompose and approximate the waveform of a deterministic signal, the SMP algorithm extends the method to approximate the *probabilistic distribution* of a random vector.

It is assumed the signal space (the spectral domain) Ω is a Hilbert space with inner product $< \cdot, \cdot >$ and induced norm $|| \cdot ||$. We also assume that the collection of all analyzing functions $D \subset \Omega$ is dense, i.e., for any $x \in \Omega$ and $\epsilon > 0$, there exist a *finite* number N of analyzing functions $w_1, w_2, \ldots, w_N \in D$ and scalars $a_1, a_2, \ldots, a_N$ such that $||x - \sum_{i=1}^{N} a_i w_i|| < \epsilon$. Without loss of generality, it is assumed $||w_j|| = 1$ for all $w_j \in D$. We call $(a_1, a_2, \ldots, a_N)$ a decomposition or *feature* of x under D. With a given distribution of the random vector x, SMP algorithm proceeds with the following steps:

1. Set $x_0 = x, j = 1$.
2. Compute $w_j = \arg\max_{w \in D} E[< x_{j-1}, w >^2]$.
3. Let $a_j =< x_{j-1}, w_j >$ and $x_j = x_{j-1} - a_j w_j$.
4. Repeat steps 2 and 3 with $j \leftarrow j + 1$ until $E[|a_j|^2] \leq \epsilon$, a predetermined threshold.

It can be verified that SMP algorithm inherits the following properties from the original matching pursuit algorithm:

- The random variables $a_1, a_2, \ldots$ are uncorrelated.
- $E[|a_j|^2]$ is monotonically decreasing with respect to j. Since $E[|a_j|^2] \geq 0$, this implies the algorithm always converges.
- The finite term realization $x_N = \sum_{j=1}^{N} a_j w_j$ converges to x in mean square sense, which, by Chebyshev inequality, implies the convergence in distribution [6].

The last property implies that the distribution of x can be approximated to any precision by the distribution of x_N. Since under $\{w_1, w_2, \ldots, w_N\}$ there exists a homeomorphic mapping between x_N and its features (as manifest by the decomposition coefficients a_j's), one may in turn approximate the distribution of x_N with that of its features. In essence, we regard the feature vector $(a_1, a_2, \ldots, a_N)$ as a reduced representation of the signal x and treat the probabilistic distribution of the feature vector as an approximation of the signal distribution. According to Chebyshev inequality, how good the approximation is can be controlled by the threshold ϵ, which naturally determines the depth of iteration N in the SMP algorithm.

3. Maximum Likelihood Training with SMP

The goal of ML training is to optimize the distribution estimation for each class of data. Let x denote the random vector modeling the signals of class C. An ML training tries to maximize the likelihood $p(x|\Phi_C)$, where Φ_C denotes the distribution parameters of class C. In our approach, we first invoke SMP algorithm for each class of data to find a class-specific analyzing functions set D_C and feature vector $(a_{C1}, a_{C2}, \ldots)$. The feature vector is then assumed to have a Gaussian mixture distribution, of which parameters Λ_C can be estimated using the expectation-maximization (EM) algorithm [3]. In essence, the parameters for each class are regarded as $\Phi_C = (D_C, \Lambda_C)$. Most importantly, the selection of D_C in this SMP/EM arrangement is now part of the training procedure and is optimized together with Λ, hence the features thus extracted will abide by the same optimality criterion as the classifier.

To be most general, assume that D_C for each class is actually composed of M clusters, i.e., $D_C = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \ldots \cup \mathcal{D}_M, \Lambda_C =$

$\lambda_1 \cup \lambda_2 \cup \ldots \cup \lambda_M$, and

$$p(x|\Phi_C) = \sum_{i=1}^{M} c_i p_i(x|\phi_i) = \sum_{i=1}^{M} c_i p_i(x|\mathcal{D}_i, \lambda_i). \quad (1)$$

Note that we have dropped the class index C for $\phi_i, \mathcal{D}_i$, and λ_i for notational simplicity. An obvious constraint $\sum_i c_i = 1$ is imposed here for $p(x|\Phi_C)$ to be a valid probabilistic measure. Let $X = \{x_k\}_{k=1}^K$ be a collection of independently observed training data for the class of interest. Therefore, the ML training is to find Φ_C that maximizes the observation likelihood $p(X|\Phi) = \prod_{k=1}^K p(x_k|\Phi)$. This can be solved by the EM algorithm. One key step in the EM algorithm is to treat each x_k as a portion of the "real" variable $y_k = (x_k, i_k)$ with i_k indicating the cluster membership of x_k, hidden from the observation. Using Jensen's inequality, it can be shown [3] that, given an initial guess of the parameters Φ', one can always improve the likelihood by finding $\Phi_C = \arg\max_\Phi E[\log p(Y|\Phi)|X, \Phi']$. From Eq. (1), we have

$$\begin{aligned} & E[\log p(Y|\Phi)|X, \Phi'] \\ = & \sum_{i_1=1}^{M} \sum_{i_2=1}^{M} \cdots \sum_{i_K=1}^{M} \\ & \left[\sum_{k=1}^{K} \log c_{i_k} p_{i_k}(x_k|\phi_{i_k}) \cdot \prod_{k=1}^{K} \frac{c'_{i_k} p_{i_k}(x_k|\phi'_{i_k})}{p(x_k|\Phi')} \right] \\ = & \sum_{i=1}^{M} \sum_{k=1}^{K} \left[\log c_i p_i(x_k|\phi_i) \cdot \frac{c'_i p_i(x_k|\phi'_i)}{p_i(x_k|\Phi')} \right] \\ = & \sum_{i=1}^{M} \log c_i \sum_{k=1}^{K} p_{ik} + \sum_{i=1}^{M} \sum_{k=1}^{K} p_{ik} \log p_i(x_k|\phi_i) \quad (2) \end{aligned}$$

where

$$p_{ik} = \frac{c'_i p_i(x_k|\phi'_i)}{p(x_k|\Phi')} \quad (3)$$

can be interpreted as the conditional probability of x_k belonging to cluster i. The optimization problem can therefore be divided into independent portions as reflected in the two separate terms in Eq. (2). The optimization of the first term, namely, c_i's, is a maximization with equality constraint problem, of which solution can be obtained through Lagrange multiplier as

$$c_i = \frac{\sum_k p_{ik}}{\sum_i \sum_k p_{ik}}.$$

Maximization of the second term involves choosing $\mathcal{D}_i$ and the corresponding distribution parameters λ_i. $\mathcal{D}_i$ can be obtained with the SMP algorithm proposed in the previous section. Note that in order to realize the distribution approximation with SMP, one must be able to compute the expectancy of the signal projected on each analyzing function (cf. step 2 in SMP). In the case of obtaining $\mathcal{D}_i$, one effectively has to know the conditional probability of each observation x_k belonging to cluster i. For this purpose, as stated before, p_{ik} in Eq. (3) serves as a good approximation. Since p_{ik} depends on Φ', the choice of $\mathcal{D}_i$ (and hence the feature extraction method) can be iteratively refined in the same manner as the model parameters. Once $\mathcal{D}_i$ is obtained, the decomposition coefficients are then regarded as the feature vector with a Gaussian mixture distribution, of which parameters λ_i can be reestimated with the conventional ML estimation method (e.g., using once again the EM algorithm).

In summary, we propose an ML training procedure in which the parameters of both the pattern matching (λ_i) and feature extraction ($\mathcal{D}_i$) are obtained based on the EM and the SMP algorithms for iterative improvement. The most important property of this SMP/EM approach is that the feature extraction process is integrated with the model optimization procedure. In contrast to many current feature extraction methods, e.g., PCA and LDA, our approach does not treat the training data as equally weighted. Rather, we effectively impose weighting on each observation by its estimated likelihood (p_{ik}) as implied in Step 2 of the SMP algorithm. Consequently, the features thus extracted are more relevant to maximize the likelihood, and hence the optimality criteria for the feature extraction and signal modeling are made consistent. During the training phase, the parameters for feature extraction and the classifier (D and Λ, respectively) are obtained together. To recognize unseen patterns x, step 3 in the SMP algorithm is first employed to compute the feature vector $(a_1, a_2, \ldots)$ for each cluster of each class, and then λ_i and Eq. (1) are applied to compute the likelihood of x for each class. This likelihood measure can then be employed as the score for the decision rule for recognition purposes.

4. Experimental Results and Discussion

Since the SMP algorithm is derived from the matching pursuit algorithm, it inherits one important caveat: the speed of convergence depends heavily upon the nature of the signals and the underlying analyzing functions w_j. For example, Mallat *et al.* demonstrates a noise removal method by choosing a set of analyzing functions that can represent signals of interest in small N but require large N to represent white noise. While in this case Mallat was able to target the signal with carefully selected analyzing functions, it is not clear what the general principles are for designing the analyzing functions. The algorithm is therefore most useful if the forms of the analyzing functions are not too complicated. For acoustic and image signals, however, one attractive candidate is from the biological systems. Recent studies suggest the auditory system seems to analyze the acoustic power spectrum with Gabor wavelets [7] (amazingly, similar Gabor tuning curves are also observed in the visual cortex [2]). With this biological motivation, we experiment the above training procedure using psychophysically calibrated Gabor functions for stationary vowel recognition.

We take the mel-scale power spectrum of the center frame of each vowel (excluding diphthongs) in TIMIT database as the observation $x(f)$, where f indicates the acoustic frequency in the mel-scale. Each analyzing function is a Gabor wavelet having the form of $w_j(f) = \rho_j \exp(-(f - f_j)^2/2\sigma_j^2) \cos(k_j(f - f_j) + \psi_j)$ with free parameters $(f_j, \sigma_j, k_j, \psi_j)$. The inner product is defined as $< x, w >= \int x(f) w(f) df$. In our experiment, we choose $M = 4$ and set the SMP stop condition at ϵ less than 5% of the signal variance. As a result, the dimensions of the feature vectors (N) range from 4 to 9. When each feature vector is modeled as a four-component Gaussian mixture, the SMP/EM algorithm achieves a

SNR	clean	30 dB	20 dB	10 dB
MFCC	65.1(77.8)	64.9(72.1)	61.2(68.5)	51.4(55.3)
SMP/EM	67.1(72.3)	66.5(72.2)	64.3(71.8)	59.6(65.0)

Table 1: Vowel recognition with additive speech-like noise. Numbers in the parentheses are the classification rates on the training data.

classification rates of 67.1% on testing set and 72.3% on training set (see Table 1). These are comparable to the same task using a representation based on 12-order mel-frequency cepstrum coefficients (MFCC) with similar complexity (16-component Gaussian mixture density models). For most vowels, there is usually a dominating cluster $\mathcal{D}_i$ with its weights (c_i in Eq. (1)) considerably larger than others (usually $c_i > 0.70$ for dominating clusters). The center frequencies (f_j) of the corresponding Gabor wavelets from the dominating cluster often coincide with the typical formant frequencies of that vowel. This may not be surprising since Gabor wavelets function much like peak detectors, and the acoustic-phonetic information of vowels is known to center around formant peaks. One issue still under investigation is whether the corresponding σ_j in the Gabor wavelet would reflect the typical formant bandwidths as clear as center frequencies.

We also compared the performance of the MFCC and SMP features under noisy conditions. We first sum up all the utterances in the TIMIT database with random gains. This results in a noise waveform that is reasonably stationary and has a speech-like spectral profile. This noise is then added to all the utterances at various levels, from which the vowel tokens are computed. For each noise level, the training and testing procedures are performed with the same parameters as described above. The results (summarized in Table 1) demonstrate a major challenge to the speech community, i.e., the performance of automatic speech recognizers usually do not have a graceful degradation in noisy environments. Part of the problems, as elaborated in the beginning of the paper, is that the commonly used feature representations (such as MFCC) are not designed to capture the speech features that can be best utilized by the classifier for pattern recognition. The joint optimization concept embedded in the SMP/EM algorithm tries to address this issue and seems to improve recognition results in this case. Its merits, however, must be further inspected with larger scale of experiments.

Like the SMP/EM algorithm proposed above, the conventional methods based on eigenvector decomposition are also designed with approximating (to the second order) the signal distribution in mind. The distinction is the conventional methods inherently assume that each independent observation is equally likely. Such an assumption is reasonable without class-specific information available at the feature extraction stage. However, our key argument in developing SMP/EM is based on the fact that one does not have to make such a limiting assumption. Most important of all, there exists a link between the classifier and feature extractor on the probabilistic measure of each signal, and such link can be gracefully incorporated into forming an integrated training paradigm.

Note that SMP/EM proposed here is a generic algorithm that can be applied not just to static patterns. One can extend this framework to hidden Markov model (HMM) based system for dynamic pattern recognition. This is plausible because EM algorithm can be nested, as demonstrated in SMP/EM itself. By embedding the Baum-Welch training algorithm for continuous density HMM, SMP/EM can therefore be included into a HMM based system in a similar fashion, treating state sequence, analyzing cluster index, and Gaussian mixture index as "hidden" information, respectively. More ambitiously, one can even include the temporal characteristics of the neural responses into the analyzing functions in the form of $w_j(t, f)$, i.e., changing the signal space from a frequency only to a time-frequency domain. With little modification, SMP/EM algorithm can be applied directly for dynamic pattern $x(t, f)$ in its time-frequency representation. Each cluster of D then corresponds to a segment rather than a frame of the signal. These issues remain active research topics.

Acknowledgment

Mr. Kuansan Wang is now with Speech Service Technologies Group, NYNEX Science and Technology Inc. The work was done when he was with Speech Research Department, AT&T Bell Laboratories.

5. REFERENCES

1. Jont B. Allen. How do human process and recognize speech? *IEEE Transactions on Speech and Audio Processing*, 2(4):567–577, October 1994.
2. Russell L. De Valois and Karen K. De Valois. *Spatial Vision*. Oxford Science Publications, New York, NY, 1990.
3. A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Annals of Royal Statistics Society*, 39:1–38, December 1977.
4. Richard O. Duda and Peter E. Hart. *Pattern classification and scene analysis*. Wiley, New York, NY, 1973.
5. Stephane Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, pages 3397–3415, December 1993.
6. Athanasios Papoulis. *Probability, random variables, and Stochastic Processes*. McGraw Hill Book Company, 1984.
7. Kuansan Wang and Shihab A. Shamma. Spectral shape analysis in the central auditory system. *IEEE Transactions on Speech and Audio Processing*, 3(5):382–395, September 1995.