A 2D EXTENDED HMM FOR SPEECH RECOGNITION

Jiayu Li

Department of Statistics
The University of Chicago
Chicago, Illinois 60637, USA

Alejandro Murua

Department of Statistics
University of Washington
Seattle, Washington 98195, USA

ABSTRACT

A two-dimensional extension of Hidden Markov Models (HMM) is introduced, aiming at improving the modeling of speech signals. The extended model (a) focuses on the conditional joint distribution of state durations given the length of utterances, rather than on state transition probabilities; (b) extends the dependency of observation densities to current, as well as neighboring states; and (c) introduces a local averaging procedure to smooth the outcome associated to transitions from successive states. A set of efficient iterative algorithms, based on segmental K-means and Iterative Conditional Modes, for the implementation of the extended model, is also presented. In applications to the recognition of segmented digits spoken over the telephone, the extended model achieved about 23% reduction in the recognition error rate, when compared to the performance of HMMs.

1. INTRODUCTION

Despite its success, the HMM framework presents three obvious limitations at the modeling level of speech production (synthesis): (i) state durations are implicitly exponentially distributed (a side-effect of the Markovian assumption); (ii) high temporal correlations in the signal are not fully captured, since data are supposed to depend only on current states; (iii) the discontinuity produced in synthesized data, when state jumps occur, is not consistent with the smoothly varying trend observed in real signals.

Researchers have acknowledged some of these shortcomings. To improve the state duration modeling, explicit state duration models, also known as semi-HMMs, have been proposed (e.g. [3]); and many extensions of HMM have appeared in the literature addressing the weak temporal correlation accounted for in the original HMM approach. Among these extensions are the frame correlated HMM [8], the conditionally Gaussian HMM [16], and two-dimensional HMMs [11]. In contrast, the continuity problem has yet to be addressed (however, the class of segment HMM [13], which is usually viewed as a method for improving temporal correlations, falls mostly within this problem).

In extensions of HMM such as the ones in [15][14], the current state as well as the previous observation are used to determine or restrict the current observation density. This approach falls short of a full solution to the temporal correlation and continuity problems (the dependency is obviously restricted to past states, in order to keep the effectiveness of dynamic programming in place), since it is more natural and precise to condition the current observation density on the preceding as well as subsequent states and corresponding observations (it is well-known that, due to the nature of the articulation process, the pronunciation of a determined phoneme is affected by both the *preceding* and *subsequent* phonemes [10, pp. 124–127]).

Two-dimensional extensions of HMMs (usually created for the analysis of spectrograms or log-spectrograms) offer a more realistic approach to speech recognition, but so far, they are not considered practical, due to a lack of efficient algorithms for their implementation (however, see [11]). In this paper, we propose a new class of two-dimensional HMMs that addresses the three limitations stated above, and that can be efficiently implemented through a set of iterative algorithms. In fact, this new class of processes: (a) model the state durations explicitly; (b) extend the dependency of observation densities to current, preceding, and subsequent states; and (c) introduce a local *smoother* that acts on mean energies associated to successive jumps of states.

We advocate two-dimensional models, because speech signal representations correspond naturally to two-dimensional objects; one of the dimensions represents time, and the other, frequency (or something equivalent, such as cepstrum coefficients, wavelet scale parameters, etc). Moreover, data are locally continuously-varying in both dimensions. We believe that, if local continuity is appropriately modeled, then global characteristics of the data will be estimated more efficiently given the same amount of training data; in fact, our experiments (see §5 below), support this belief.

Since this new approach gives up the single directed temporal dependence relationship that prevails in the usual HMMs framework, no parallel to the Viterbi, and Baum-Welch algorithms that fit HMMs can be applied here (dynamic programming is not effective in this new set-up). Instead, to solve the estimation problem, a procedure based on the segmental K-means algorithm [7], is proposed; and to compute fast estimates of probabilities, an iterative procedure similar to the Iterative Conditional Modes algorithm (ICM) [1] is suggested (see §4).

We have applied our model to a speaker-independent recognition problem, involving segmented digits, spoken over the telephone (see §5); when compared to the performance of HMMs with about the same number of parameters, on the same task, our algorithm achieved about 23%

reduction in the error rate.

This paper is organized as follows. Section 2 describes the basic time-frequency representation used by our models. Section 3 introduces our two-dimensional extension of HMMs, and discusses how these new processes overcome the three limitations stated above. Section 4 deals with the algorithms involved in the estimation of parameters. Section 5 contains the results related to the particular application of our model to the recognition of segmented (spoken) digits.

2. SPEECH SIGNAL REPRESENTATION

We consider the normalized log-spectrogram associated to each utterance of a word, as the basic object representing the speech signal. The frequency domain corresponds to the output of a bank-of-filters signal processing front-end. In our studies, we work with the Bark frequency scale, as described in [10, pp. 159–161]. Normalization is achieved by averaging all frequency band energies over the entire time-frequency span of each utterance; this yields a two-dimensional array of observations $\{O_{tf} : t = 1, \ldots, T,\ f = 1, \ldots, F\}$, where T stands for the length or duration (measured in overlapping frames) of the utterance, and F, for the number of frequency bands (about 20 or less, depending on the task).

For fixed $\tau \geq 1$ and $\phi \geq 1$, we define a neighborhood structure over the lattice $\mathcal{L} = \{1, \ldots, T\} \times \{1, \ldots, F\}$, much in the same way as it is done over Gibbs random fields [4]. Associated to each lattice point $(t, f) \in \mathcal{L}$ there are two neighborhoods composed of nearest-neighbor lattice points, denoted and given by $\partial(t, f) = \{(t', f') : \max\{1, t - \tau\} \leq t' \leq \min\{T, t + \tau\},\ \max\{1, f - \phi\} \leq f' \leq \min\{F, f + \phi\},\ (t', f') \neq (t, f)\}$ (a *first order* neighborhood); and $\partial^2(t, f) = \{(t'', f'') : (t'', f'') \in \partial(t', f') \text{ for } (t', f') \in \partial(t, f)\}$ (a *second order* neighborhood). These neighborhoods play central roles in the characterization of our two-dimensional model.

3. A TWO-DIMENSIONAL HMM

Our model for each unit of speech, e.g. word, phoneme, or syllable, consists of a double stochastic array $\{Y_{tf}, X_t : t = 1, \ldots, T,\ f = 1, \ldots, F\}$, where $\{Y_{tf}\}$ is the *observation process,* and $\{X_t\}$ is the *state process.*

3.1. The State Process

The process $\{X_t\}$ takes values on a finite and ordered state space $\mathcal{S} = \{q_1 \leq q_2 \leq \cdots \leq q_r\}$, and is assumed to be a left-to-right process, i.e. $X_t \leq X_s$ if and only if $t \leq s$; it replaces the Markov chain process of the usual HMM framework. A crucial departure from this latter framework is our interest on the joint distribution of $\{X_t\}$ given the length T of the utterance, rather than on state transition probabilities. We assume that this joint distribution depends only on the visited states and the associated lengths of the visits. More precisely, let the random variable D_i = length of visit to state q_i, $i = 1, \ldots, r$; then $P(X_1 = s_1, \ldots X_T = s_T|T) = P(D_1 = d_1, \ldots, D_r = d_r|T)$, where d_i = number of times $s_t = q_i$, $i = 1, \ldots, r$, $s_1, \ldots, s_T \in \mathcal{S}$. D_i is assumed to be a mixture of a point mass $m_0(q_i)$ at zero, and a truncated Poisson with mean rate λ_i, $i = 1, \ldots, r$; this assumption yields

$$P(s_1, \ldots, s_T|T) = \prod_{i=1}^{r} \{(1 - m_0(q_i))\,(1 - \delta_{0d_i}) \times e^{-\lambda_i}(1 - e^{-\lambda_i})^{-1}(\lambda_i^{d_i}/d_i!) + m_0(q_i)\delta_{0d_i}\}$$

where $\delta_{0d_i} = 1$ if and only if $d_i = 0$, and it is zero otherwise.

3.2. The Observation Process

The dependency of each observation variable Y_{tf} on the whole set of observations $\{Y_{t'f'}\}$ excluding itself, given the state process, is restricted to its neighbors (see §2) $Y_{\partial(t,f)} = \{Y_{t'f'} : (t', f') \in \partial(t, f)\}$. In what follows, $O_{-(tf)}$ will stand for the set of all observations in the lattice $\mathcal{L}$ excluding O_{tf}, $(t, f) \in \mathcal{L}$.

Let s_1^T denote a realization of the random vector $\{X_t\}$, and $s_{\partial(t,f)}$, $s_{\partial^2(t,f)}$, denote those state values associated to the first and second order nearest neighbors of (t, f), respectively (notice that, by an abuse of notation, $s_{(t',f')}$ refers to $s_{t'}$); we write

$$P(Y_{tf} = O_{tf}|O_{-(tf)}, s_1^T) = P(O_{tf}|O_{\partial(t,f)}, s_{\partial^2(t,f)}) \quad (1)$$

i.e., we assume that the density of Y_{tf} not only depends on the current state $X_t = s_t$, but also on neighboring observations and states. Our model incorporates in this way the high temporal correlation observed in speech signals. We further assume that (1) corresponds to a Gaussian density with mean $\mu_{tf} + \Sigma_{(t',f')\in\partial(t,f)}\{c_{t'-t,f'-f}\ (O_{t'f'} - \mu_{t'f'})\}$, and variance σ_{tf}^2; where $\mu_{tf} = E\{Y_{tf}|\ X_{\partial(t,f)}\}$, and $\sigma_{tf}^2 = \mathrm{Var}\,\{Y_{tf}|Y_{\partial(t,f)}, X_t\} = \sigma^2(X_t, f)$ (here, $E\{\cdot|\cdot\}$ and $\mathrm{Var}\,\{\cdot|\cdot\}$ denote conditional expectation and conditional variance, respectively; and $X_{\partial(t,f)}$ is the analog of $s_{\partial(t,f)}$). Thus, the mean energy μ_{tf} at time t and frequency f, not only depends on the current state, but also on the preceding and subsequent states, $s_{\partial(t,f)}$. Detailed parameterization of μ_{tf} is given below. σ_{tf}^2 is the conditional variance given the neighboring states as well as neighboring observations. It is assumed to depend on the state sequence only through the current state (in our experiments with spoken digits –see below–, σ_{tf}^2's do not present strong variations over the time-frequency domain). The *projection coefficient* $c_{t'-t,f'-f}$ accounts for interaction between observations; it measures the amount of the "reflection" (see [9, Chapter 5] for details) from observation $O_{t'f'}$ onto O_{tf}; it is assumed to be invariant to both time and frequency translations. The projection coefficients are expected to be high, since high positive correlation between energies on neighboring lattice points is usually present over the time-frequency domain. We note that the special case in which $c_{t'-t,f'-f} = 0$, for all $(t, f), (t', f') \in \mathcal{L}$, and μ_{tf} depends only on the current state, corresponds to an ordinary HMM.

Symmetry constraints impose the well-known "detailed balance" relation $c_{t'-t,f'-f}\ \sigma_{t'f'}^2 = c_{t-t',f-f'}\ \sigma_{tf}^2$. Without any special requirement on the σ_{tf}^2's, the detailed balance relation will not be satisfied. However, the relation will be approximately satisfied if σ_{tf}^2's do not vary too much (a fact observed in our experiments; see §5), and the condition $c_{t'-t,f'-f} = c_{t-t',f-f'}$ is imposed.

It can be shown [9] that given a state sequence s_1^T, $\{Y_{tf}\}$ is a Gaussian random field (GRF). A sufficient condition [9, Chapter 5] to ensure that the variance-covariance matrix associated to this field be positive-definite, is to constrain the projection coefficients to be small, namely

$$\sum_{(t',f')\in\partial(t,f)} |c_{t'-t,f'-f}| < 1 \qquad (2)$$

In our applications, we further constrain these coefficients to be positive. This is justified by the positive correlation observed among adjacent frequency energies.

Dealing with the spectral continuity problem. In HMMs, μ_{tf} depends only on the current state s_t. As we mentioned earlier, artificial discontinuities in the modeling of the time-frequency dynamics of speech signals are produced when state jumps occur. A possible solution, which is the one implemented in our model, is to allow μ_{tf} to not only depend on the current state, but also on neighboring preceding and subsequent states. More explicitly, we assume that $\mu_{tf} = E\{Y_{tf}|X_1^T\} = \mu(X_t, f) + \sum_{(t',f')\in\partial(t,f)} a_{t'-t,f'-f}\mu(X_{t'}, f')$, where $\mu(s,f)$, the *state-dependent mean energy,* depends only on the current state s and frequency f; it corresponds to the usual mean energy in HMMs. For simplicity, the *smoothing coefficients* $\{a_{t'-t,f'-f}\}$ are assumed to be invariant to translations in time, and frequency. The $(2\tau+1)\times(2\phi+1)$ matrix $\{a_{t'-t,f'-f}\}$ corresponds to a local *smoother* on the state-dependent mean energy vector $\{\mu(s_t,f)\}$. Its role is to smooth the outcome of possible jumps of states.

Generalization to Gaussian Mixtures. In incorporating Gaussian mixtures into the model, we follow the approach of Juang and Rabiner in [6], on partitioned Gaussian mixtures, since it simplifies both implementation and computations in our algorithms. In fact, within this approach, the choice of a mixture component $m_t \in \{1,\ldots,M\}$ (M is the maximum number of mixtures allowed in a state), as accounting for observation O_t, can be viewed as a visit to a sub-state of state s_t (which now can be thought of as a super-state); this sub-state can be denoted as (s_t, m_t). With this observation in mind, we can replace the state sequence s_1^T by the state-mixture sequence (s_1^T, m_1^T), in all the above considerations, and proceed exactly as we did before. In fact, it can be easily seen that given the state-mixture sequence, the observed process $\{Y_{tf}\}$ is again a GRF. In order to simplify the computations, we assume that given a state sequence s_1^T, $P(m_1,\ldots,m_T|s_1^T) = \prod_{t=1}^T P(m_t|s_t)$, i.e. the occurrence of a mixture component m_t at state s_t is independent of the occurrence of any other mixture component $m_{t'}$ at any other time $t' \neq t$, given the state sequence s_1^T. In what follows we will denote $P(m_t|s_t)$ by γ_{m_t,s_t}.

4. PARAMETER ESTIMATION

The set of parameters determining our model, which we will denote by Λ, is given by $\Lambda = (\{\mu(q_i,m,f)\}, \{\sigma^2(q_i,f)\}, \{c_{dt,df}\}, \{a_{dt,df}\}, \{\gamma_{m,q_i}\}, \{m_0(q_i)\}, \{\lambda_i\})$, where $q_i \in \mathcal{S}$, $i=1,\ldots,r$, $f=1,\ldots,F$, $dt=-\tau,\ldots,\tau$, $df=-\phi,\ldots,\phi$, $m=1,\ldots,M$ (notice that $\sigma^2(q_i,m,f)=\sigma^2(q_i,f)$, is assumed to be constant over all mixture components associated to state q_i, $i=1,\ldots,r$).

Due to the intractability of the quantity

$$P(O_1^T|\Lambda) = \sum_{s_1^T,m_1^T} P(O_1^T|(s_1^T,m_1^T),\Lambda)P(m_1^T|s_1^T,\Lambda)P(s_1^T|\Lambda)$$

under our model assumptions, we estimate $P(O_1^T|\Lambda)$ by $\widetilde{P}(O_1^T|\Lambda)$, which is given by

$$\max_{s_1^T,m_1^T} P(O_1^T|(s_1^T,m_1^T),\Lambda)P(m_1^T|s_1^T,\Lambda)P(s_1^T|\Lambda) \qquad (3)$$

We refer to this latter quantity as the *maximum likelihood path probability.* A justification for this estimate, is that in practical applications, the term associated to the optimal state sequence, is the term that contributes almost all the weight in the above sum.

Our training procedure maximizes $\widetilde{P}(O_1^T|\Lambda)$ over the parameter space Λ. The algorithm is based on the segmental K-means algorithm (see for example [7]), and iterates between the following two steps: **(A)** the first step is to fix the state-mixture sequence and maximize $P(O_1^T|\Lambda)$ over Λ. We note that this maximization corresponds to a constrained optimization problem, since the projection coefficients $\{c_{dt,df}\}$ must be positive and satisfy condition (2). The optimization can be done by separately maximizing the three factors in (3), since the sets with parameters involved in each of the factors are mutually disjoint. As a result of this factorization, explicit expressions for the optimal $\{\gamma_{m,q_i}\}$, $\{\lambda_i\}$ and $\{m_0(q_i)\}$ are easily obtained [9]:

$$\hat{\gamma}_{m,q_i} = \frac{\text{number of visits to mixture component } m}{\text{length of visit to state } q_i}$$

$$\hat{\lambda}_i = \text{average length of visits to state } q_i$$

$$\hat{m}_0(q_i) = \text{proportion of utterances that skipped state } q_i$$

The remainder parameters are estimated using a gradient descent technique over the parameter space. **(B)** the second step consists of finding the optimal state-mixture sequence for fixed values of Λ, i.e. it consists of finding $\widetilde{P}(O_1^T|\Lambda)$. This step solves the so-called decoding or time-alignment problem, and is crucial for recognition purposes (see §5). We propose an iterative procedure to get estimates of the optimal state-mixture sequence. In each iteration, an orderly sweep is done over the state-mixture sequence, updating each state s_t and mixture component m_t so as to maximize $\widetilde{P}(O_1^T|\Lambda)$ given that the remainder states and mixture components in the sequence remain fixed. In our experiments described below, this algorithm converged very fast, requiring about two or three iterations. This procedure carries the same idea proposed by Besag [1] in his Iterative Conditional Modes (ICM) algorithm, to compute a maximum a posteriori (MAP) estimate of Markov random fields parameters.

5. EXPERIMENTS

We applied our model to the recognition of segmented "digits" *(one, two, ... ,nine, zero* and *oh),* spoken over the telephone. The data were taken from the CSLU Number Corpus of the Center for Spoken Language Understanding of the Oregon Graduate Institute of Science and Technology.

This corpus is a real world application that contains "fluent numbers" spoken by thousands of people when saying numbers such as their street address numbers, zip-codes, and telephone numbers. False starts, pauses, repetition, and background noise are very common in these data, and make the task difficult (see [2] for more details). The corpus is divided into three sets of 8829, 3052, and 3119 speech files; the first one is reserved for training, the second one, for development, and the last one, for testing. We located and worked with all occurrences of the eleven digits in the corpus.

The number of states r in each digit model was determined by the particular phoneme configuration of the corresponding digit. However, for simplicity, we imposed the same number of mixture components ($M = 4$) in each state of the eleven models. τ and ϕ were set to one.

Recognition. Given a test sample O_1^T, our procedure recognizes it as an utterance of a determined digit d, if $\Lambda_d = \arg\max_{\Lambda_{d'}} P(\Lambda_{d'}|O_1^T) = \arg\max_{\Lambda_{d'}} P(O_1^T|\Lambda_{d'})$ (here we assume a uniform prior on word frequencies), where $\Lambda_{d'}$ stands for the model corresponding to digit d'. As explained in §4, we estimate $P(O_1^T|\Lambda_d)$ by $\widetilde{P}(O_1^T|\Lambda_d)$, which is computed through an algorithm similar to ICM.

Results and Conclusions. Since this task can be regarded as an isolated speech recognition task with a fairly small vocabulary, we decided to jointly train the eleven digit models using the minimum classification error (MCE) criterion. In fact, it has been argued [5] that MCE methods are superior to maximum likelihood estimation methods when the assumed distributions given by the models, do not correspond to the true ones. We observed this fact in our experiments, as well (see [9] for more details).

Model parameters were estimated with an algorithm very similar to the one described in §4. Several values of the projection coefficients were tried on all eleven models, in order to explore the parameter space; those parameters that minimized the classification error over a validation set (corresponding to a subset of the training corpus), were chosen to test our models. The overall test error rate was 10.7%, which is considered small for this task, given that previous studies involving a much larger number of model parameters [12][17] reported error rates of about 5%–12% on similar tasks involving the same database. For comparison purposes, we also fitted HMMs to the same task (this corresponds to setting the projection and smoothing coefficients to zero). Our model yielded a 23% reduction in the error rate, when compared to the performance of HMMs. In fact, the solely incorporation of the smoothing coefficients $\{a_{dt,dt}\}$ (setting the projection coefficients to zero) yielded a reduction of about 6% in the error rate. It is worth-noting that we simply model the log-spectrogram of the words, and do not introduce any other features (e.g. cepstrum coefficients, power differences) in the observation process; hence, the improvement in recognition rates achieved by our model, is due only to a more realistic modeling of the speech signal process, and not to a better extraction of feature vectors.

6. REFERENCES

[1] J. Besag. On the statistical analysis of dirty pictures. *J. Roy. Stat. Soc., Ser. B*, 48(3):259–302, 1986.

[2] R. A. Cole, M. Fanty, and T. Lander. Telephone speech corpus development at CSLU. In *Proc. Int. Conf. Spoken Language Processing*, 1994.

[3] J. D. Ferguson. Hidden Markov analysis: An introduction, 1980. Institute for Defense Analysis, Princeton, New Jersey.

[4] D. Geman and S. Geman. Stochastic relaxation, Gibbs distributions and the bayesian restoration of images. *IEEE-PAMI*, 6:721–741, 1984.

[5] B. H. Juang, W. Chou, and C. H. Lee. Minimum classification error rate methods for speech recognition. *IEEE Trans. Speech, Audio Processing*, 5(3):257–265, 1997.

[6] B. H. Juang and L. R. Rabiner. Mixture autoregressive hidden Markov models for speech signals. *IEEE Trans. Acoust., Speech, Signal Processing*, 33(6):1404–1413, 1985.

[7] B. H. Juang and L. R. Rabiner. The segmental K-means algorithm for estimating parameters of hidden Markov models. *IEEE Trans. Acoust., Speech, Signal Processing*, 38(9):1639–1641, 1990.

[8] N. S. Kim and C. K. Un. Frame-correlated hidden Markov model based on extended logarithmic pool. *IEEE Trans. Speech, Audio Processing*, 5(2):149–160, 1997.

[9] J. Li. *Modeling Log-Spectrograms For Applications To Speech Recognition Using HMM and Gaussian Random Fields.* PhD thesis, Department of Statistics, University of Chicago, Chicago, Illinois, 1998.

[10] P. Lieberman and S. Blumstein. *Speech Physiology, Speech Perception, and Acoustic Phonetics.* Cambridge University Press, 1988.

[11] H. Lucke. Improved acoustic modeling for speech recognition using 2D Markov random fields. In *Proc. ICASSP-95*, pages 540–543, 1995.

[12] K. W. Ma. Applying large vocabulary hybrid HMM-MLP methods to telephone recognition of digits and natural numbers. Technical Report 95-024, International Computer Science Institute, Berkeley, California, 1995.

[13] M. Ostendorf, V. V. Digalakis, and A. O. Kimball. From HMM's to segment models: A unified view of stochastic modeling for speech recognition. *IEEE Trans. Speech, Audio Processing*, 4(5):360–378, 1996.

[14] K. K. Paliwal. Use of temporal correlation between successive frames in hidden Markov model based speech recognizer. In *Proc. ICASSP-93*, pages 215–218, 1993.

[15] S. Takahashi. Phoneme HMM's constrained by frame correlations. In *Proc. ICASSP-93*, pages 219–222, 1993.

[16] C. J. Wellekens. Explicit time correlation in hidden markov models for automatic speech recognition. In *Proc. ICASSP-87*, pages 384–386, 1987.

[17] Y. Yan, M. Fanty, and R. Cole. Speech recognition using neural networks with forward-backward probability generated targets. In *Proc. ICASSP-97*, 1997.