Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 61 (2015) 328 – 333

Procedia Computer Science

Complex Adaptive Systems, Publication 5
Cihan H. Dagli, Editor in Chief
Conference Organized by Missouri University of Science and Technology
2015-San Jose, CA

Autoregressive Hidden Markov Model and the Speech Signal

Jacob D. Bryan*, Stephen E. Levinson†

University of Illinois at Urbana-Champaign, Champaign IL 61801, United States

Abstract

This paper introduces an autoregressive hidden Markov model (HMM) and demonstrates its application to the speech signal. In this variant of the HMM the observed signal is assumed to be Gaussian autoregressive and the probability density function is derived based on an approximation of the linear prediction error. A Baum-Welch style set of re-estimation formulas are then derived and used to infer the model parameters for a given data set, which correspond to linguistic structure in the context of speech data. The new set of re-estimation formulas are then applied to speech data and experimental results demonstrate inference of broad phonetic categories without prior knowledge of linguistic information. The experimental results and stability of this model are then briefly contrasted with historic experiments wherein phonetic information has been inferred directly from the speech signal using a similar autoregressive model.

Keywords: hidden Markov model; autoregressive HMM; linear prediction; speech signal processing

1. Introduction

In this paper, we seek to develop a method of inferring linguistic structure from the speech signal in an unsupervised manner. Given that the hidden Markov model (HMM) has been demonstrated to infer linguistic structure from text[1], we accomplish this task by applying an HMM directly to the speech. Previously, the autoregressive HMM or hidden

* *E-mail address:* jdbryan2@illinois.edu
† *E-mail address:* slevins@ifp.illinois.edu

doi:10.1016/j.procs.2015.09.151

filter model has been shown to infer broach phonetic categories from the speech signal. In the case studied by Poritz, linear prediction analysis was incorporated into the HMM and a set of Baum-Welch style re-estimation formulas were developed[2,3]. This approach was based on the covariance method of estimating linear prediction coefficients (LPCs)[4]. In order to extend these results, an alternative method of performing linear prediction analysis, namely the autocorrelation method, is incorporated into the HMM so that the inferred filter parameters carry a guarantee of stability. The resulting re-estimation formulas are then applied to speech data and broad phonetic and phonotactic information is inferred.

In the remainder of this paper, we will develop this new autoregressive HMM and its corresponding Baum-Welch algorithm. This model will then be applied to a set of speech data and the results will be presented. Finally, we will contrast this new model with the Poritz model and suggest future applications in speech processing.

2. Model Development

The problem of inferring linguistic structure directly from the speech signal may be broken into two parts. Namely, we must model the linguistic structure within the signal as well as the spectral behavior of the signal at a given moment in time. In the case of this work, we will use the HMM to model linguistic structure and the all-pole filter model of speech to capture the spectral distribution. We will then incorporate the all-pole filter model into the observation probability of the HMM so as to infer the filter parameter alongside the HMM parameters.

2.1. Hidden Markov Model

The HMM is a mathematical model that can be used to infer sequential patterns in a set of observations by assuming that a given sequence of observations is generated by an unobserved sequence of states. The elements of the observation sequence are denoted by O_t and the elements of the state sequence are denoted by s_t. In order to simplify the mathematics, the transitions between states are assumed to have the Markov property, in that the probability of state s_t taking a particular value is only dependent on the preceding state, s_{t-1}. In addition, we assume that the probability of a given observation at any point in the sequence is only dependent on the underlying state.

Using this structure, we can characterize the system with a set of state transition probabilities $a_{ij} = Pr\{\text{transition from state } i \text{ to state } j\}$ and state dependent observation probabilities $b_j\,(O_t) = Pr\{O_t\,|s_t = j\}$. In the case that the observations come from a discrete and finite set V of size M, we can characterize the conditional probability of observing each element of that set as $b_{jk} = b_j\,(O_t = v_k)$ where $v_k \in V$. Based on these definitions, an HMM with n states can be characterized by a set of parameters $\boldsymbol{\lambda} = \{\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}\}$, where $\boldsymbol{A} = [a_{ij}]$ is the $N \times N$ state transition matrix, $B = [b_{jk}]$ is the $N \times M$ observation matrix, and $\pi = (1,0,\ldots,0)$ is the $1 \times N$ initial state vector. Using a set of parameters $\boldsymbol{\lambda}$, the forward-backward algorithm as shown in (1), (2) and (3) can be used to efficiently compute the probability of an observation sequence of length T. The forward and backward probabilities are initialized as $\alpha_1(j) = \pi_j b_j(O_t)$ and $\beta_T(j) = 1\ \forall j$ respectively.

$$\Pr\{O|\boldsymbol{\lambda}\} = \sum_{j=1}^{N} \alpha_T(j) \tag{1}$$

$$\alpha_{t+1}(j) = \left[\sum_{j=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_t), \qquad 1 \leq t \leq T-1 \tag{2}$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(O_t) \beta_{t+1}(j)\,, \qquad T-1 \geq t \geq 1 \tag{3}$$

This in turn leads to the Baum-Welch algorithm given by the set of re-estimation formulas (4) and (5), where the over-bar denotes a new estimate of that parameter.

$$\overline{a}_{ij} = \frac{\sum_{t=1}^{T-1} \alpha_t(i) a_{ij} b_j(O_t) \beta_{t+1}(j)}{\sum_{t=1}^{T-1} \alpha_t(i)\beta_t(i)} \tag{4}$$

$$\overline{b}_{jk} = \frac{\sum_{t \ni O_t = v_k} \alpha_t(j)\beta_t(j)}{\sum_{t=1}^{T-1} \alpha_t(j)\beta_t(j)} \tag{5}$$

These formulas represent a contraction map such that the optimal values occur at a fixed point within the parameter manifold. The optimal parameter values are then determined by iterating the Baum-Welch algorithm until such a fixed point is reached[3,5,6,7].

2.2. All-pole Filter Model of Speech

The speech signal is often used as a canonical example of a nonstationary signal. Fortunately, the spectrum of speech is largely dependent on the configuration of the vocal tract at any given point in time. Since the shape of the vocal tract changes at a rate much slower than the frequencies of the speech signal, we may assume that changes to the spectrum of the signal are negligible over a sufficiently short period of time. In practice, the short-time stationary assumption is applied for analysis windows up to approximately 30 ms in duration[4].

Under this assumption, we may model the vocal tract as an all pole filter as shown in (6) where the gain is denoted by σ and the filter coefficients, or LPCs, are denoted by c_k.

$$H(z) = \frac{\sigma}{1 - \sum_{k=1}^{p} c_k z^{-k}} \tag{6}$$

Using this model, we obtain the time domain representation of the signal $s[n]$, shown in the synthesis equation (7) where $e[n]$ represents the excitation signal. We can then estimate the model parameters by solving a system of at least p such equations, requiring at least $2p$ samples of the signal. Alternatively, the filter coefficients may be estimated from the autocorrelation function of the signal using the Yule-Walker equation, shown in (8) where $r[n]$ is the autocorrelation function of lag n. Although the autocorrelation method generally leads to a larger error than the direct approach, the Toeplitz structure of the autocorrelation matrix guarantees that the resulting filter will be stable[4].

$$s[n] = \sum_{k=1}^{p} c_k s[n-k] + \sigma e[n] \tag{7}$$

$$\tilde{\boldsymbol{c}} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{bmatrix} = \begin{bmatrix} r[0] & r[1] & \cdots & r[p-1] \\ r[1] & r[0] & \cdots & r[p-2] \\ \vdots & \vdots & \ddots & \vdots \\ r[p-1] & r[p-2] & \cdots & r[0] \end{bmatrix} = \boldsymbol{R}^{-1}\boldsymbol{r} \tag{8}$$

2.3. Parameter Estimation

In its original construction, Poritz modelled the observation probability as a Gaussian autoregressive process, which is derived by assuming that the excitation signal $e[n]$ is Gaussian[2]. Alternatively, we can approximate the error over the analysis window at time t using the linear prediction residual as given by Itakura[8]. This approach yields the observation probability given by (9) where $\boldsymbol{c}_j = \left[1, -\tilde{\boldsymbol{c}}_j^{\,T}\right]^T$ is the set of LPCs corresponding state j.

$$b_j(O_t) = \frac{2}{\sqrt{2\pi}\sigma_j} \exp\left(\frac{\boldsymbol{c}_j^T \boldsymbol{R}_t \boldsymbol{c}_j}{2\sigma_j^2}\right) \tag{9}$$

Using the observation model (9), a set of Baum-Welch style re-estimation formulas may be developed via the procedure outlined by Baum et. al.[7] or by applying the Expectation Maximization algorithm[5]. The resulting formulas are shown in (10), (11) and (12), where $\boldsymbol{R}_j$ and $\boldsymbol{r}_j$ are the autocorrelation matrix and vector corresponding to state j as defined by the weighted sum of autocorrelation functions shown in (10). It is significant to note that since $\boldsymbol{R}_j$ is a weighted sum of Toeplitz matrices, its symmetry is preserved and so is the stability of the resulting LPCs. In addition, the Toeplitz structure of the matrices allows for the LPCs to be efficiently computed using the Levinson-Durbin recursion[4]. Lastly, it should be noted that since the state transition probabilities are independent of the observation model, their re-estimation formulas remain as shown in (4). These re-estimation formulas form a contraction map that be recursively applied to estimate filter parameters and state transition probabilities until the model parameters converge to a fixed point[5].

$$r_j[n] = \sum_{t=1}^{T} \alpha_t(j)\beta_t(j)r_t[n] \tag{10}$$

$$\bar{\sigma}_j^2 = \frac{\boldsymbol{c}_j^T \boldsymbol{R}_j \boldsymbol{c}_j}{\sum_{t=1}^{T} \alpha_t(j)\beta_t(j)} \tag{11}$$

$$\bar{\boldsymbol{c}}_j = \left[\,1, -\boldsymbol{R}_j^{-1}\boldsymbol{r}_j\right]^T \tag{12}$$

3. Application to Speech

Since this observation model is derived from the formula posed by Itakura, this method is equivalent choosing filter parameters that minimize the LPC distance. Although this distance is not a true metric, its utility in speech recognition has been well established[8,9,10]. With this in mind, the re-estimation formula were applied to several recordings of speech data taken from the TIMIT Acoustic-Phonetic Continuous Speech Corpus[11].

The set of recordings from a single speaker were collected to make up 30 seconds of speech data sampled at a rate of 8 kHz. The data was then segmented into overlapping analysis windows using a 30 ms Hamming window at 5 ms step sizes. The autocorrelation function was then computed for each windowed segment of the signal. This sequence of autocorrelation functions r_t, was then used as input to the autoregressive HMM. The Baum-Welch algorithm was then iterated until the updated estimates for all model parameters were within 10^{-4} of the previous estimates. In addition, the state transition probabilities were smoothed using the Good-Turing estimate after each iteration of the algorithm so as to prevent the state transition probabilities from converging to zero incorrectly[5,12].

3.1. Experimental Results

An autoregressive HMM with 5 internal states and 5 filter coefficients was applied to the given speech data to produce the following results. In order to more clearly interpret the results, the probability of each state at each time step is computed by (13) and shown in Figure 1. In addition, the linguistic interpretation of each state was obtained via playback by grouping together segments of the data corresponding to a high probability of that state. This information is listed alongside the state transition probabilities in Table 1. Lastly, the frequency response of each of the state dependent filters is shown in Figure 2. As is evident in these results, the inferred states correspond to broad phonetic categories and the state probabilities give an indication of the phonotactic structure in the speech signal.

$$\gamma_j(t) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \tag{13}$$

These results are closely related to those produced by Poritz[2], however key differences should be noted. First, the filter model used in this experiment contains a larger number of coefficients than the Poritz model. This is due to the fact that Poritz's method was derived from a more direct estimation of the LPCs. Consequently, the state dependent filters were estimated more accurately but without a guarantee of stability which causes the model to be sensitive to initial conditions. Instead, this new model is able to estimate a set of stable filters and the margin of error may be arbitrarily reduced by simply adding more filter coefficients. As a result, this new model produces much more stable results and is able to consistently converge from a randomized initialization, provided that the model order of the state dependent filters is sufficiently high.

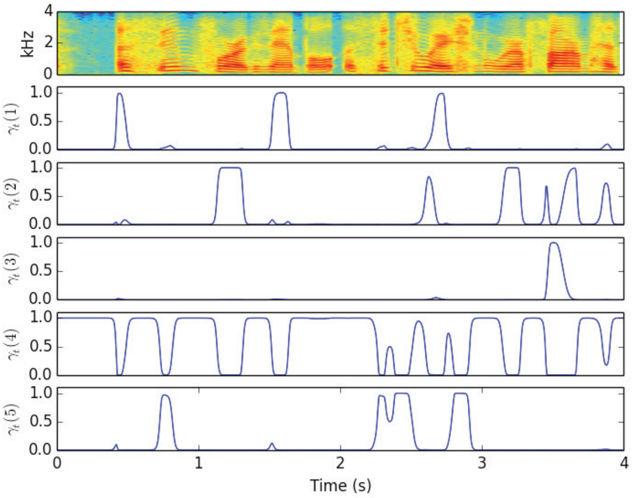
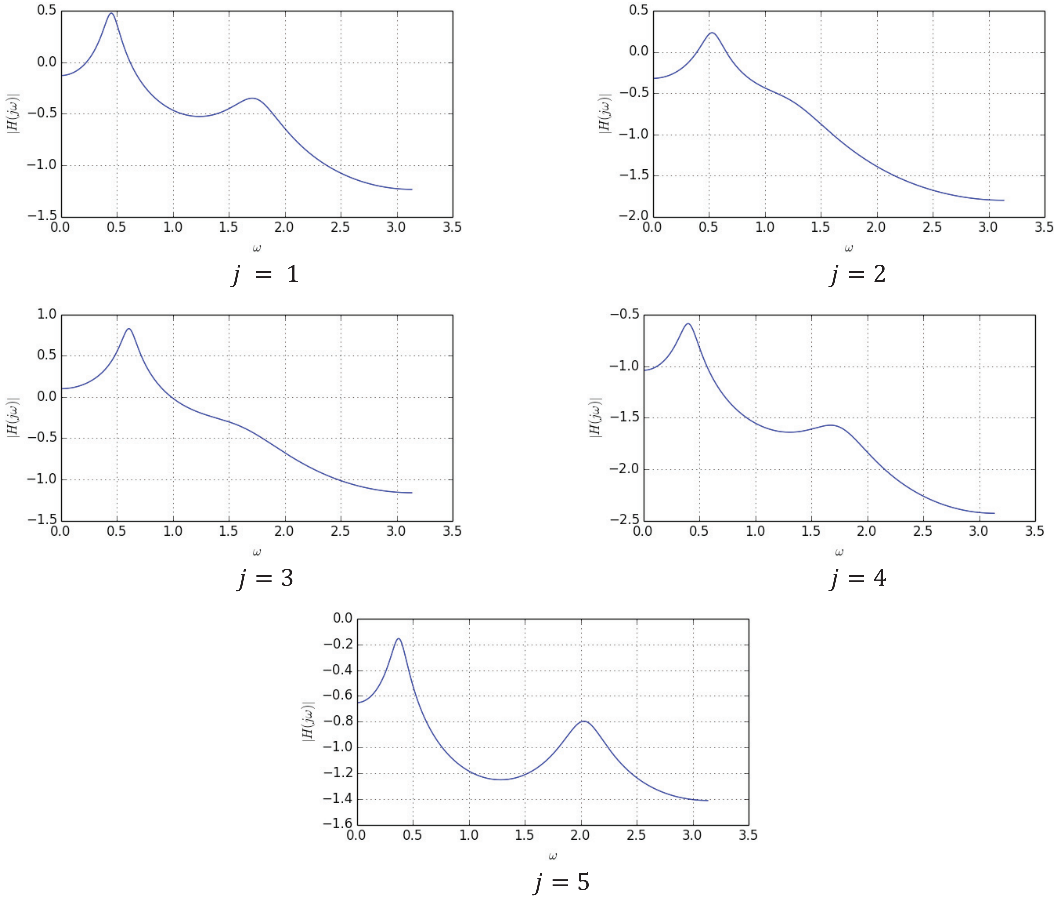


Figure 1: Spectrogram of the speech signal and probability histories of each state.



$j = 1$

$j = 2$

$j = 3$

$j = 4$

$j = 5$

Figure 2: Frequency response of the inferred state dependent all-pole filters

Table 1: Inferred state transition matrix and linguistic interpretation of each state.

From \To	1	2	3	4	5	Interpretation
1	0.930	0.013	0.000	0.056	0.000	Vowels (eh)
2	0.007	0.944	0.008	0.040	0.000	Semivowels
3	0.018	0.033	0.948	0.000	0.000	Vowels (ah)
4	0.002	0.011	0.000	0.978	0.009	Plosives and Silence
5	0.020	0.000	0.000	0.036	0.944	Fricatives

4. Conclusion

In this paper, we have presented a variant of the autoregressive hidden Markov model that is developed based on the autoregressive method of linear prediction. A set of re-estimation formulas were developed and this model was demonstrated to produce stable estimates of the all-pole filter model of the speech signal. The resulting Baum-Welch algorithm was then applied to a set of speech data and the broad phonetic categories of the data were inferred. Given the stability of this algorithm, this model represents a method of inferring the linguistic structure of the speech signal in an unsupervised manner. It is suggested that this model might be used for the purpose of detecting word or syllable boundaries or as a first stage of signal processing for speech recognition. Lastly, because the underlying filters within this model are stable, this method may also prove useful for the purpose of speech synthesis.

References

1. R. L. Cave and L. P. Neuwirth, “Hidden Markov models for English,” in *Proc. Symp. on the Application of Hidden Markov Models to Text and Speech*, J. D. Ferguson, Ed., Princeton NJ, pp. 16-56, 1980.
2. A. B. Poritz, “Linear predictive hidden Markov models and the speech signal,” in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 1291-1294, 1982.
3. A. B. Poritz, “Hidden Markov models: A guided tour,” in *Acoustics, Speec, and Signal Processing, International Conference on*, IEEE, pp. 7-13, 1988.
4. T. F. Quatieri, *Discrete-Time Speech Signal Processing*, Ser. Prentice-Hall signal processing series. Upper Saddle River, NJ: Prentice Hall PTR 2002.
5. S. E. Levinson, *Mathematical Models for Speech Technology*, ser. Prentice-Hall signal processing series. The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, England: John Wiley and Sons Ltd, 2005.
6. L. E. Baum, “An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process,” *Inequalities*, vol. 3, pp. 1-8, 1970.
7. L.E. Baum, T. Petrie, G. Soules, and N. Weiss, “A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains,” *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164-171, 1970.
8. F. Itakura, “Minimum prediction residual principle applied to speech recognition,” *Acoustics, Speech and Signal Processing*, *IEEE Transactions on*, vol. 23, no. 1, pp. 67-72, 1975.
9. H. Sakoe and S. Chiba, “Dynamic Programming Algorithm Optimization for Spoken Word Recognition,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 1, pp. 43-49, 1978.
10. L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson, “Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition,” *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 6, pp. 575-582, 1978.
11. J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallet, and N. Dahlgren, “TIMIT acoustic-phonetic continuous speech corpus LDC93S1,” *Philadelphia: Linguistics Data Consortium*, 1993.
12. I. J. Good, “The population frequencies of species and population parameters,” *Biometrika*, vol. 40, no 3/4, pp. 237-264, 1953.