MULTIVARIATE-STATE HIDDEN MARKOV MODELS FOR SIMULTANEOUS TRANSCRIPTION OF PHONES AND FORMANTS

Mark Hasegawa-Johnson

Department of ECE, University of Illinois
Urbana, IL 61801, USA

ABSTRACT

A multivariate-state HMM — an HMM with a vector state variable — can be used to find jointly optimal phonetic and formant transcriptions of an utterance. The complexity of searching a multivariate state space using the Baum-Welch algorithm is substantial, but may be significantly reduced if the formant frequencies are assumed to be conditionally independent given knowledge of the phone. Operating with a known phonetic transcription, the multivariate-state model can provide a maximum *a posteriori* formant trajectory, complete with confidence limits on each of the formant frequency measurements. The model can also be used as a phonetic classifier by adding the probabilities of all possible formant trajectories. A test system is described which requires only nine trainable parameters per formant per phonetic state: five parameters to model formant transitions, and four to model spectral observations. Further simplifications were achieved through parameter tying.

1. INTRODUCTION

This article proposes an algorithm which simultaneously transcribes the phonetic content and the formant frequencies of an utterance. Previous attempts to use formant frequencies in speech recognition require the formant tracker to make a hard decision about the frequency of each formant before the commencement of phonetic classification. In contrast, the algorithm proposed in this article seeks phonetic and formant transcriptions which describe the spectral content of the utterance in a jointly optimal manner. In effect, the algorithm proposed in this article is a soft-decision formant tracker.

The structure of the formant tracker is based loosely on the HMM formant tracker proposed by Kopec [3]. In Kopec's HMM formant tracker, the state of the hidden Markov model is a vector of formant frequencies $\vec{\phi}_t = [\phi_{1,t}, \ldots, \phi_{P,t}]$, where P is the number of formants to be tracked. Kopec reduced the complexity of his formant tracker by vector quantizing $\vec{\phi}_t$, and performing a Viterbi search in the space of vector quantization indices, in order to determine the maximum likely formant frequency track $\Phi^* = [\vec{\phi}^*_1, \ldots, \vec{\phi}^*_T]$. Once calculated using Kopec's formant tracker, the matrix Φ^* was then appended to the observation vector of an HMM digit recognizer [1]. The observation vector also included the LPC spectrum from which the formants were originally calculated, so Bush and Kopec note that the formant frequencies were effectively a repetition of information already available to the recognizer. Even so, the second formant slope $d\phi^*_{2,t}/dt$ was found to increase digit recognition scores; information about other formants had no significant effect on recognition performance.

In this article, Kopec's formant frequency state vector is augmented with the scalar state variable q_t of a traditional HMM:

$$\vec{q}_t \equiv [q_t, \phi_{1,t}, \ldots, \phi_{P,t}] \qquad (1)$$

Section 2 describes an efficient Viterbi search, and an approximate Baum-Welch algorithm, in which the cost of searching the state space given in equation 1 is significantly reduced using appropriate independence assumptions. Section 3 describes a simple test system which can be used to track formant frequencies at the release of a voiced stop consonant, or to identify the consonant. Section 4 describes parameter tying and estimation, and section 5 gives results.

2. SEARCH IN A MULTIVARIATE STATE SPACE

The formant frequencies may be modeled as a vector of hidden state variables which influence but do not determine the acoustic spectrum. In the notation of speech recognition, the scalar state variable q_t in traditional HMM formulations [4] is replaced by the vector state variable $\vec{q}_t$ in equation 1. Assume that the scalar phonetic variable q_t takes on N_q different values, while each of the formant frequencies $\phi_{i,t}$ takes on N_ϕ different values; then the vector state variable $\vec{q}_t$ takes on a total of $N_q N_\phi^P$ different values.

Maximum *a posteriori* transcription of formant frequencies and phonetic information given a spectrogram can be accomplished using either the Viterbi algorithm or the Baum-Welch algorithm. This article will derive a computationally efficient multivariate approximation of the Baum-Welch algorithm. A computationally efficient multivariate Viterbi algorithm may be derived in much the same way.

The Baum-Welch algorithm calculates the probability $\gamma_t(\vec{j}) = P(O, \vec{q}_t = \vec{j}|\lambda)$, where $O = [\vec{o}_1, \ldots, \vec{o}_T]$ is the matrix of observation vectors, and λ is the model. $\gamma_t(\vec{j})$ is the product of a forward probability $\alpha_t(\vec{j})$ and a backward probability $\beta_t(\vec{j})$, which are calculated using recursions of the form:

$$\alpha_t(\vec{j}) = p(\vec{o}_t|\vec{j}) \sum_{\vec{i}} \left[\; \alpha_{t-1}(\vec{i}) \; p(\vec{j}|\vec{i}) \; \right] \qquad (2)$$

This work was part of the author's doctoral thesis [2]. The work was performed under the supervision of K. Stevens at MIT, with grant support from the NIH.

In the general case, the only way to solve equation 2 is by summing over all of the possible state vectors. The total computation required, with $N_q N_\phi^P$ possible state vectors and T time steps, is $T(N_q N_\phi^P)^2$.

It is possible to construct a simplified approximation of equation 2 by assuming that formant frequencies $\phi_{p,t}$ are dependent on the "phonetic state" q_t (which is equivalent to the state in a traditional HMM), but that given knowledge that q_t equals soem particular state j_0 and $q_{t-1} = i_0$, the formant frequencies $\phi_{p,t} = j_p$ and $\phi_{r,t} = j_r$ are independent:

$$p(\vec{j}|\vec{i}) = p(j_0|i_0) \prod_{p=1}^{P} p(j_p|i_0, i_p) \tag{3}$$

$$p(\vec{o}_t|\vec{j}) \propto \prod_{p=1}^{P} p(\vec{o}_t|j_0, j_p) \tag{4}$$

If the transition and observation probabilities satisfy equations 3 and 4, then equation 2 can be approximated using the equations

$$\alpha_t(\vec{j}) \approx \tilde{\alpha}_t(\vec{j}) = \prod_{p=1}^{P} \alpha_t(j_p, j_0) \tag{5}$$

$$\alpha_t(j_p, j_0) = p(\vec{o}_t|j_p, j_0) \times \tag{6}$$

$$\sum_{i_0} \sum_{i_p} \left[\; \alpha_{t-1}(i_p, i_0) \; p(j_0|i_0)^{1/P} \; p(j_p|i_p, i_0) \; \right]$$

The total complexity of finding $\tilde{\alpha}_t(\vec{j})$ using equation 6 for T time steps is $TP(N_q N_\phi)^2$.

Even if the transition and observation probabilities satisfy equations 3 and 4, equation 6 is only an approximation of equation 2. It can be shown that a Viterbi search can be simplified in a similar manner without suffering a similar loss of accuracy; unfortunately, a Viterbi search does not provide us with the probability $\gamma_t(\vec{j})$. It can also be shown that $\tilde{\alpha}_t(\vec{j}) \geq \alpha_t(\vec{j})$, with equality if the phonetic state q_τ is known with certainty for $\tau < t$.

The multivariate-state algorithm may be used as a maximum *a posteriori* formant tracker. The *a posteriori* distribution of each of the formant frequencies given observations $\vec{o}_1, \ldots, \vec{o}_T$, and given the phoneme sequence $q_1, \ldots, q_T$, is:

$$\gamma_t(j_p|q_t) = \alpha_t(j_p|q_t)\beta_t(j_p|q_t) \tag{7}$$

$$\alpha_t(j_p|q_t) = p(\vec{o}_t|j_p, q_t) \times \tag{8}$$

$$\sum_{i_p} \left[\; \alpha_{t-1}(i_p|q_{t-1}) \; p(j_p|i_p, q_{t-1}) \; \right]$$

The maximum *a posteriori* (MAP) formant transcription of an utterance is the sequence of formants $\phi_{p,t}^*$ which maximize $\gamma_t(\phi_{p,t}|q_t)$.

It may be the case that an application needs to know how confident the formant tracker is about the estimated formant frequency $\phi_{p,t}^*$. Confidence limits for the formant $\phi_{p,t}^*$ are given by the distribution $\gamma_t(\phi_{p,t}|q_t)$. For example, it is possible to numerically integrate $\gamma_t(\phi_{p,t}|q_t)$, and to find two frequencies f_1 and f_2 such that the true value of $\phi_{p,t}$ is between f_1 and f_2 with a probability of 95%.

Phoneme classification with the multivariate-state model can be accomplished using either the Viterbi algorithm or the approximate Baum-Welch algorithm. Using the Baum-Welch algorithm, maximum-likelihood phone classification is accomplished by choosing the phone model λ which maximizes

$$P(O|\lambda) = \sum_j \alpha_T(\vec{j}) \tag{9}$$

where T is the length of the waveform. As noted previously, exact computation of equation 9 requires on the order of $T(N_q N_\phi^P)^2$ computations, which is usually not practically feasible. An approximate maximum likelihood classification can be accomplished by choosing λ to maximize

$$\tilde{P}(O|\lambda) \equiv \prod_{p=1}^{P} \sum_{j_0} \sum_{j_p} \alpha_T(j_p, j_0) \tag{10}$$

It can be shown that $\tilde{P}(O|\lambda) \geq P(O|\lambda)$, with equality if the sequence $q_1, \ldots, q_T$ is deterministic.

3. TEST SYSTEM

The approximate Baum-Welch algorithm described in equations 5 through 10 was tested using stop consonant releases excised by hand from the TIMIT database. Formant tracking and phonetic classification were performed using segments consisting of six consecutive non-overlapping 10ms frames, beginning 5ms before stop release.

Each of the three places of articulation (lips, tongue blade, and tongue body) was modeled using two hidden Markov models: one for male speakers, and one for female speakers. Each model consisted of six states, with no self-loops and no skipped states; in other words, given knowledge of the model, the state sequence $[q_1, \ldots, q_6]$ was deterministic and non-repeating.

Formant frequencies tend to be slowly varying, so it is reasonable to model $p(j_p|i_p, i_0)$ using a unimodal distribution, with a peak somewhere near $j_p = i_p$. The test system modeled formant transitions using a two-dimensional Gaussian distribution with phoneme-dependent parameters:

$$p\left([i_p, j_p]|i_0\right) = \mathcal{N}\left([i_p, j_p], \bar{\phi}_p(i_0), \Sigma_{\phi,p}(i_0)\right) \tag{11}$$

where $\mathcal{N}(\vec{x}, \bar{x}, \Sigma)$ is the normal distribution with mean vector $\bar{x}$ and covariance matrix Σ. Since j_p is a discrete random variable, the PDF in equation 11 was converted to a PMF using approximate numerical integration.

The dependence of the observed DFT spectrum on any given formant frequency j_p was modeled using a strictly local, independent modulation of the relative amplitude $A(j_p, t)$ and spectral convexity $C(j_p, t)$ at the specified frequency j_p, with probability densities whose parameters depend only on the phonetic state j_0:

$$p(\vec{o}_t|j_p, j_0) = p(A(j_p, t)|j_0)p(C(j_p, t)|j_0) \tag{12}$$

The spectral amplitude $A(j_p, t)$ is defined as the ratio of the squared DFT spectrum $|X(j_p, t)|^2$ to the total DFT energy $E(t)$, where

$$E(t) = \sum_{k=0}^{N/2} |X(kF_s/N, t)|^2, \tag{13}$$

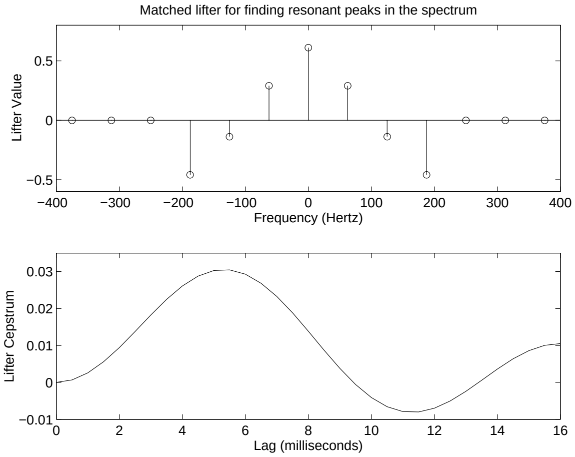

Figure 1: Lifter used to estimate spectral convexity.

and N is the number of samples in the DFT spectrum, and F_s is the sampling rate. Since ratio variables are not well modeled by a Gaussian distribution, the ratio is transformed using a form of the logistic transform:

$$A(j_p, t) \equiv 10 \log_{10} \left(\frac{|X(j_p, t)|^2}{E(t) - |X(j_p, t)|^2} \right) \tag{14}$$

The resulting amplitude measure $A(j_p, t)$ is modeled using a Gaussian distribution, with phone-dependent mean and variance:

$$p(A(j_p, t)|j_0) = \mathcal{N}\left(A(j_p, t), \bar{A}_p(j_0), \sigma^2_{A,p}(j_0)\right) \tag{15}$$

The convexity measurement $C(j_p, t)$ is an approximation of the second derivative of the log spectrum, created by convolving $\log|X(f,t)|$ with a seven-sample FIR lifter:

$$C(j_p, t) = 20 \sum_{k=-3}^{3} H(k) \log_{10} |X(j_p - k, t)| \tag{16}$$

The lifter $H(k)$ was designed by truncating the spectrum of a 125Hz-bandwidth complex pole pair, and normalizing the result to zero mean and unit energy. The spectrum and cepstrum of $H(k)$ are shown in figure 1.

The probability distribution of $C(j_p, t)$ is not well modeled by a Gaussian distribution. When human judges measure formant frequencies, they usually locate a formant frequency near a local maximum of $C(j_p, t)$, implying that larger values of $C(j_p, t)$ are somehow "better" than average values — in other words, the mode of $p(C(j_p, t))$ should be higher than the mean. Since the true form of $p(C(j_p, t))$ was unknown, convexity was modeled using a "half-Gaussian" weighting distribution (not a true probability distribution), in which all convexity values above the training sample mean were considered equally likely:

$$p(C(j_p, t)|j_0) = \mathcal{N}\left(\underline{C}(j_p, t), \bar{C}_p(j_0), \sigma^2_{C,p}(j_0)\right) \tag{17}$$

$$\underline{C}(j_p, t) \equiv \min\left(C(j_p, t), \bar{C}_p(j_0)\right) \tag{18}$$

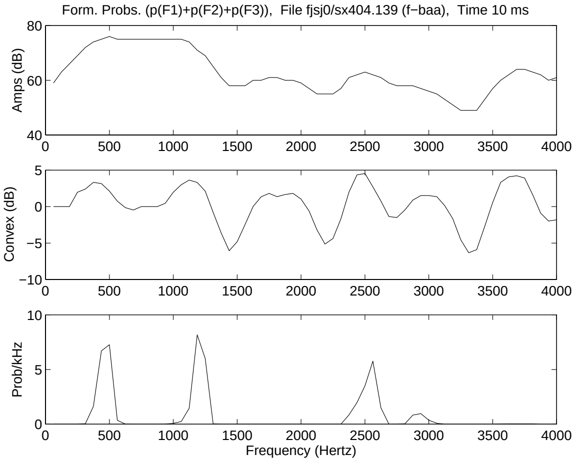

Figure 2: Spectrum, spectral convexity, and *a posteriori* formant probabilities, as calculated by the formant tracker, 10ms after release of the first /b/ in "Barb." *A posteriori* probability distributions of all three formants are shown on the same plot, because there is no significant overlap between the distributions.

4. PARAMETER TYING AND ESTIMATION

The probability densities in equations 11, 15, and 17 are specified by nine parameters per formant per phonetic state: five parameters to model the formant transition (equation 11), and four to model spectral observations (equations 15 and 17). Visual analysis of the training data suggests that amplitude and convexity parameters vary little from one place of articulation to another and from one speaker gender to another, so the amplitude mean and variance parameters were tied across place of articulation and speaker gender. The parameters of equation 11, on the other hand, vary substantially as a function of place and gender, but are more slowly varying as a function of time. The parameters $\Sigma_{\phi,p}(j_0)$ were therefore tied together across all states within each HMM, i.e. $\Sigma_{\phi,p}(q_t) = \Sigma_{\phi,p}(\lambda)$, where λ denotes place of articulation and gender of the speaker. After these simplifications, the average number of trainable parameters per formant per phonetic state is 3.17. Further tying reduced the complexity to 1.84 trainable parameters per formant per phonetic state; see [2] for details.

The multivariate-state model was tested in two experiments, with separate training for each experiment.

In the first experiment, the model was tested as a soft-decision formant tracking algorithm, using the *a posteriori* formant probability distributions given in equation 7. The model parameters were estimated using manual transcriptions of the formant trajectories following 36 stop release tokens, produced by 36 different speakers (18 male, 18 female). Training tokens were manually segmented, and the mean and variance of each Gaussian distribution in equations 11 and 12 were estimated using sample means from each segment, pooled across training tokens.

In the second experiment, the model was used to classify the place of articulation of stop releases. For this experiment, the model was trained using 3141 non-word-final

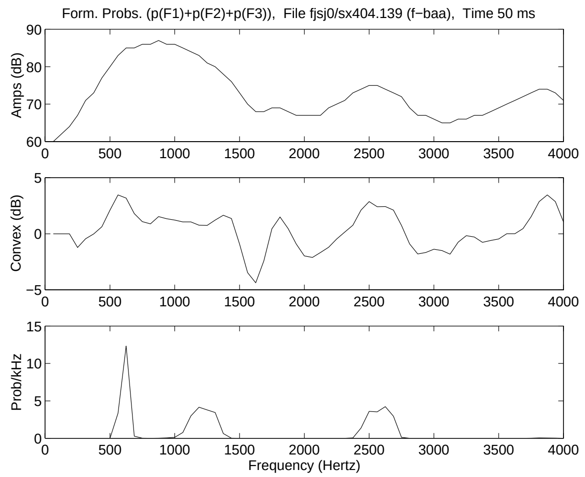

Figure 3: Spectrum, spectral convexity, and *a posteriori* formant probabilities, as calculated by the formant tracker, 50ms after release of the first /b/ in "Barb."

Vowel	Correct	Num.	Classified As:		
	Phone	Toks.	/b/	/d/	/g/
Low Back	/b/	63	92		
/aa,	/d/	29		83	
ah/	/g/	64		17	81
High Front	/b/	216	94		
/iy,y,	/d/	102		76	14
ih,ux/	/g/	33		15	79
High Back	/b/	18	83		11
/uw,w,	/d/	25		88	12
uh/	/g/	31			84
Other	/b/	474	90		
/eh,ae,	/d/	476	15	72	13
ix,ax/	/g/	474		11	81

Table 1: Stochastic formant model classification of voiced stops, as a function of right context. Entries of 10% or less have been omitted.

voiced stop release tokens extracted from the TRAIN subdirectory of TIMIT. Segmentation was based on the TIMIT segmentation, but was checked manually for each training token. Formant frequencies were transcribed by the ESPS formant tracker (Entropic Research Labs, Washington, DC); formants missed by the ESPS tracker were filled in using an expectation maximization algorithm.

5. RESULTS

The multivariate-state model can be used as an MAP formant tracker by calculating, at each time t, the set of formant frequencies which maximize the *a posteriori* probability distribution $\gamma_t(\phi_p|q_t)$. If the model is used in this way, $\gamma_t(\phi_p|q_t)$ itself can be viewed as a frame-by-frame estimate of the measurement uncertainty of the formant-tracking algorithm.

Figures 2 and 3 show the spectrum, spectral convexity, and *a posteriori* formant probabilities $\gamma_t(\phi_p|q_t)$ at $t = 10$ms and $t = 50$ms following the /b/ release in the word "Barb" spoken by a female speaker. In most cases, there is a clear peak in the spectrum near the expected value of formant ϕ_p, and as a result, $\gamma_t(\phi_p|q_t)$ has low variance. In figure 3, however, the F1 peak obscures the location of the F2 peak, and as a consequence, $\gamma_t(\phi_2|q_t)$ in figure 3 is more diffuse. A different kind of uncertainty is visible in the F3 region in figure 2: there are two clear peaks in the spectrum. The algorithm is unable to definitively rule out either peak, so $\gamma_t(\phi_3|q_t)$ is non-zero for candidate values of ϕ_3 in both possible formant locations.

The model was also tested in a phoneme classification task. The approximate Baum-Welch algorithm described in equation 10 was used to classify the place of articulation of 2005 non-word-final voiced stop releases in vowel, schwa, and glide right contexts extracted from the TEST subdirectory of TIMIT. Table 5 lists confusion matrices for four different categories of right context; on average, classification of the stop consonant was 83% correct.

6. CONCLUSIONS

This article describes a hidden Markov model which simultaneously tracks the phonetic state of an utterance and three hidden formant frequencies. The model can be used as a maximum *a posteriori* formant tracker, in which case the *a posteriori* distributions $\gamma_t(j_p|j_0)$ provide a frame-by-frame estimate of formant measurement uncertainty. The model can also be used for phoneme classification or recognition.

The observation PDF used in this article models only one sample of the spectral amplitude and one sample of the spectral convexity per formant per phonetic state. The limited observations may explain the rather poor classification performance of the test system. If a more complete observation spectrum can be devised which approximately satisfies the independence constraint in equation 4, the classification performance of the model is expected to improve.

REFERENCES

[1] Marcia A. Bush and Gary E. Kopec. Network-based connected digit recognition. *IEEE Trans. Acoust., Speech, and Sig. Proc.*, ASSP-35(10):1401–1413, October 1987.

[2] Mark Hasegawa-Johnson. *Formant and Burst Spectral Measurements with Quantitative Error Models for Speech Sound Classification.* PhD thesis, MIT, Cambridge, MA, August 1996.

[3] Gary E. Kopec. Formant tracking using hidden markov models and vector quantization. *IEEE Trans. Acoust. Speech Sig. Proc.*, 34(4):709–729, Aug. 1986.

[4] Lawrence Rabiner and Bing-Hwang Juang. *Fundamentals of Speech Recognition.* Prentice-Hall, Englewood Cliffs, NJ, 1993.