MULTIBAND AUDIO MODELING FOR SINGLE-CHANNEL ACOUSTIC SOURCE SEPARATION

Manuel J. Reyes-Gomez[1], *Daniel P. W. Ellis*[1], *Nebojsa Jojic*[2]

[1]LabROSA, Dept. of Electrical Engineering, Columbia University
[2]Microsoft Research

ABSTRACT

Detailed hidden Markov models (HMMs) that capture the constraints implicit in a particular sound can be used to estimate obscured or corrupted portions from partial observations, the situation encountered when trying to identify multiple, overlapping sounds. However when the complexity and variability of the sounds are high, as in a particular speaker's voice, a detailed model might require several thousand states to cover the full range of different short-term spectra with adequate resolution. To address the tractability problems of such large models, we break the source signals into multiple frequency bands, and build separate but coupled HMMs for each band, requiring many fewer states per model. Modeling each frequency band independently, as in multiband speech models proposed by the ASR community, will result in many non-natural full spectral states. To prevent this and to enforce consistency within and between bands, at any given frame the state in a particular band is determined by the previous state in that band and the states in the adjacent bands. Coupling the bands in this manner results in a grid like model for the full spectrum. Since exact inference of such a model is intractable, we derive an efficient approximation based on variational methods. Results in source separation of combined signals modeled with this approach outperform the separation obtained by full-band models.

1. INTRODUCTION

Detailed hidden Markov models (HMMs) of audio signals can be used to separate acoustic mixtures between the sources by searching for combinations of state sequences that give the greatest agreement with combined observations. Good separation, however, requires detailed source models that might require several thousand full spectra states. In [1], HMMs with 8000 states of dimension 513 are used, which constitute a total of 8000 × 513 $= 41,040,000$ spectral parameters in the model. Such a large number of parameters presents many challenges during both learning and inference. If we break the full spectra representation in different synchronized bands (fig. 1 a), and then use separate HMMs in each band with many fewer states, we could represent a large number of full spectral configurations with substantially fewer parameters. For instance if we divide a 513 dimension full spectrum into 19 equal bands of dimension 27, we can potentially represent $19^{27} = 3.36e + 34$ full spectrum states, using only 15,390 spectral parameters. But if we train each band model independently without any constraints between bands (as in the multiband speech models used in [2] and [3]), we will obtain for many frames unnatural combinations of subband states that are not representative of the speaker. To prevent this and to enforce consistency within and between bands, we couple adjacent bands in such a way that the state in each band is determined by the previous states in that band as well as the two adjacent bands. Coupling the bands in this manner results in a grid-like model for the full spectrum. Exact inference of such a model is intractable, but we have derived an efficient approximation based on variational methods [4]. We build speaker models using several arrangements of the proposed structure, varying the number of bands and number of spectral coefficients per band, using both linear and perceptual (Bark or logarithmic) scales. We then use these models to separate mixtures of signals of the the modeled speakers. Our approach outperforms in time, efficiency, and separation quality, the performance in the same task of full-spectrum HMMs trained on the same speakers. The results also show the improvement in the separation results obtained by the band coupling, when compared to the results obtained with band models trained independently.

2. INFERENCE AND LEARNING IN THE MULTIBAND MODEL

A Hidden Markov Model is represented as a graphical model in fig. 2a, where the hidden variables $S = (s_1, s_2, .., s_T)$ represent the unknown states of the model at any given frame, and the observed variables $X = (x_1, x_2, ..., x_T)$ represent the observed feature vectors. The joint probability of the model is given by $P(X, S) = \prod_{t=1}^{T} p(s_t \mid s_{t-1}) p(x_t \mid s_t)$.

Inference and learning is performed by the EM algorithm which reaches the local maximum of the log-likelihood of the model, $L(X, \theta)$, by iteratively optimizing a bound on the log-likelihood $\mathcal{L}(X, Q, \theta)$ with respect to an auxiliary function $Q(S)$ and the model's parameters θ (the transition matrix, μ_j and C, for $p(x_t \mid s_t = j) = \mathcal{N}(x_t, \mu_j, C)$). The bound in the log-likelihood $\mathcal{L}(X, Q, \theta)$ is defined through the Jensen Inequality and for an HMM has the form:

$$L(X, \theta) \geq \mathcal{L}(Q, \theta) = \sum_S Q(S) \log \frac{P(X, S \mid \theta)}{Q(S)} \quad (1)$$

$$\begin{aligned} &= \sum_S \sum_t Q(S)(\log p(s_t \mid s_{t-1}) + \log p(x_t \mid s_t)) \\ &\quad - \sum_S Q(S) \log Q(S)) \end{aligned} \quad (2)$$

It is well known that if $Q(S)$ is found without any restrictions, the value that maximizes $\mathcal{L}(X, Q, \theta)$ such that $\mathcal{L}(X, Q, \theta) = L(X, \theta)$ is $Q(S) = P(S \mid X)$ the posterior of the hidden variables given the observations, which is

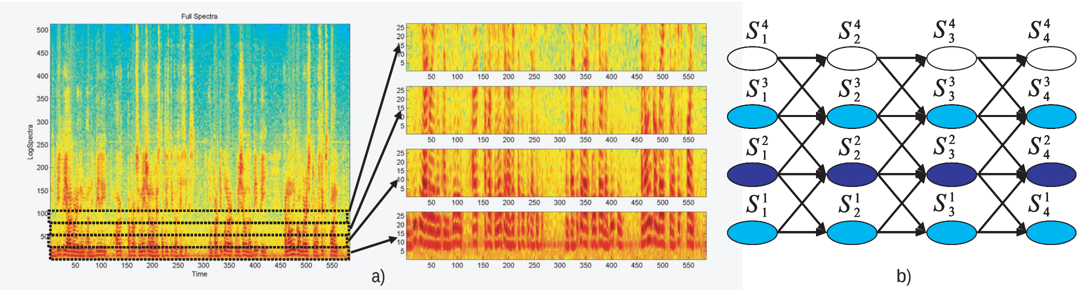
Full Spectra
LogSpectra
Time
a)
S_1^4 S_2^4 S_3^4 S_4^4
S_1^3 S_2^3 S_3^3 S_4^3
S_1^2 S_2^2 S_3^2 S_4^2
S_1^1 S_2^1 S_3^1 S_4^1
b)

Fig. 1. a) Spectrogram Partition and b) multiband model.

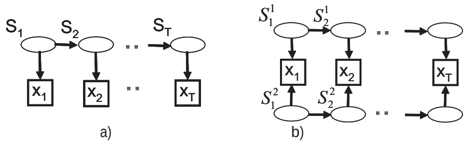

Fig. 2. a) Standard HMM and b) Factorial HMM.

found for equations with the form of eqn. (2), by the well-known forward/backward algorithm[5]. The proposed multiband model (fig. 1b), has in turn the hidden variables $S = (s_1^1, s_2^1, .., s_T^1, s_1^2, s_2^2, .., s_T^2, ...s_1^K, s_2^K, .., s_T^K)$ and the observation variables $X = (x_1^1, x_2^1, .., x_T^1, x_1^2, x_2^2, .., x_T^2, ...x_1^K, x_2^K, .., x_T^K)$ (not shown on the figure) where s_t^k and x_t^k, represent the state and the observation at time t at frequency band k.

The joint probability for this model is given by:

$$P(S,X) = \prod_{k,t} p(s_t^k \mid s_{t-1}^k, s_{t-1}^{k-1}, s_{t-1}^{k+1}) \prod_{k,t} p(x_t^k \mid s_t^k) \quad (3)$$

Its parameters are defined for each band k by the transition probabilities $p(s_t^k \mid s_{t-1}^k, s_{t-1}^{k-1}, s_{t-1}^{k+1})$, the means μ_j^k and variance C, for $p(x_t^k \mid s_t^k = j) = \mathcal{N}(x_t^k, \mu_j^k, C^k)$. For this model the bound in the log-likelihood is given by:

$$\mathcal{L}(Q,\theta) = \sum_S Q(S) \sum_{k,t} (log(p(s_t^k \mid s_{t-1}^k, s_{t-1}^{k-1}, s_{t-1}^{k+1})) + log(p(x_t^k \mid s_t^k))) - \sum_S Q(S) log(Q(S)) \quad (4)$$

Optimizing eqn. (4) with respect to $Q(S)$ with $S = (s_1^1, s_2^1, .., s_T^1, s_1^2, s_2^2, .., s_T^2, ...s_1^K, s_2^K, .., s_T^K)$ is intractable due to the large variable space. However, if we restrict the auxiliary function to be of the form $Q(S) = \prod_k (Q(S^k))$ where $S^k = (s_1^k, s_2^k, .., s_T^k)$ are the hidden variables in the HMM for the k_{th} band, the bound in the log-likelihood with this particular $Q(S)$ becomes:

$$\mathcal{L}(Q,\theta) = \sum_k \sum_{S^k,t} Q(S^k) \Big(log(p(s_t^k \mid s_{t-1}^k, s_{t-1}^{k-1}, s_{t-1}^{k+1})) + log(p(x_t^k \mid s_t^k)) \Big) - \sum_{S^k,k} Q(S^k) log(Q(S^k)) \quad (5)$$

Since we are restricting the form that $Q(S)$ can take, we can not perform any more exact inference, since $P(S \mid X)$ could never have the restricted form. However we could still do variational inference by optimizing eqn. (5) with respect to the $Q(S^k)$ factors[6], which are now called variational parameters.

The restricted auxiliary function actually decouples the hidden variables between different subband HMMs in the sense that we can optimize the set of variational parameters $Q(S^k)$ for the k^{th} band without optimizing the variational parameters for the other bands, treating them in fact as constant or "observed". We will see below, however, that the coupling between bands is actually not lost. So for any given band (k) we can express $\mathcal{L}(Q,\theta)$ as a the summation of two elements, $\mathcal{L}_k(Q,\theta)$ which contains all the variational and model parameters correspondent to the band k, and $\mathcal{L}_{\neq k}(Q,\theta)$ which does not contains any parameters for that band, and with $\mathcal{L}(Q,\theta) = \mathcal{L}_k(Q,\theta) + \mathcal{L}_{\neq k}(Q,\theta)$. Optimizing the $\mathcal{L}(Q,\theta)$ with respect to the parameters for band k is equivalent to optimizing $\mathcal{L}_k(Q,\theta)$ with respect of the same parameters. Approximating $p(s_t^k \mid s_{t-1}^k, s_{t-1}^{k-1}, s_{t-1}^{k+1})$ by
$p(s_t^k \mid s_{t-1}^k, s_{t-1}^{k-1})\ p(s_t^k \mid s_{t-1}^k, s_{t-1}^{k+1})$, we find that $\mathcal{L}_k(Q,\theta)$ takes the form:

$$\mathcal{L}_k(Q,\theta) = \sum_{S^k} \sum_t Q(S^k)(\log(\tilde{p}(s_t^k \mid s_{t-1}^k)) + [\log(\tilde{p}(s_{t-1}^{k-1}, s_{t-1}^{k+1} \mid s_t^k))] + \log(p(x_t^k \mid s_t^k))) + H(Q(S^k)) \quad (6)$$

$$\log(\tilde{p}(s_t^k \mid s_{t-1}^k)) = \sum_{S^{k-1}} Q(S^{k-1}) \log p(s_t^k \mid s_{t-1}^k, s_{t-1}^{k-1}) + \sum_{S^{k+1}} Q(S^{k+1}) \log p(s_t^k \mid s_{t-1}^k, s_{t-1}^{k+1}) \quad (7)$$

$$\log(\tilde{p}(s_{t-1}^{k-1}, s_{t-1}^{k+1} \mid s_t^k)) = \sum_{S^{k-1}} Q(S^{k-1}) \log p(s_t^{k-1} \mid s_{t-1}^{k-1}, s_{t-1}^k) + \sum_{S^{k+1}} Q(S^{k+1}) \log p(s_t^{k+1} \mid s_{t-1}^{k+1}, s_{t-1}^k) \quad (8)$$

where eqns. (7) and (8) are called the expected log-transition probability and the expected log-state likelihood given the "observed" adjacent bands. (For the full mathematical derivation please refer to [7].) In these terms is evident the actual coupling between adjacent bands models which prevents the formation of unnatural combinations of subband states by enforcing consistency within and between bands. The former term can be seen as the "weighted" log-transition matrix for the HMM for the k^{th} band. While the latter is more similar to the observation log-likelihood term $(\log(p(x_t^k \mid s_t^k)))$, here the likelihood of a given state s_t^k on the k^{th} band does not depend on the "observed" feature vector, but rather in the "observed" dynamic behavior on the adjacent subbands on the same and previous frames.

Equation (6) resembles eqn. (2), the bound in the log-likelihood for a single HMM, but with the addition of the term in the bracket, which as we have mentioned is the state log-likelihood given the "observed" adjacent bands. Then we can also apply the forward/backward algorithm, using eqn. (7) as the transition matrix and adding eqn. (8) to the observation likelihood. The parameters at each band are learned using standard M-Step [7]. Summarizing, the bands are coupled through equations (7) and (8), and then each band HMM is trained as in the case of a single isolated HMM.

3. TRAINING OF THE MULTIBAND MODEL

Each band's HMM is trained as a single HMM once eqns. (7) and (8) have been calculated. To calculate these terms, however, we need to know or have 'observed' the variational parameters of the adjacent bands. For instance if we want to train the HMM of the darkest nodes in fig. 1b, we need to know the variational parameters for the lightly-shaded nodes. Once we finish the training of the dark band, we proceed up the model to train the next band up, using the variational parameters of the just-trained band. We continue this procedure until we reach the highest band; a complete pass from the lowest to the highest is called an iteration of the multiband model. For the first iteration, all the variational parameters are uniformly initialized, meaning that the the upper band (only) in each HMM training in the first iteration is being very poorly approximated n as uniform. We continue this process until the bound stops increasing. For pseudo code for this procedure refer to ([7]).

4. SOURCE SEPARATION USING DETAILED LOG-SPECTRA MODELS

Detailed log-spectral models of speech can be used to separate combined speech signals using the "refiltering" and "log-max" technique introduced in [1]. The idea behind this approach is that when two clean speech signals are mixed additively in the time domain, the log-spectrogram of the mixture is almost exactly the maximum of the individual log-spectrograms [8], i.e. given speech signals $x_1(t)$ and $x_2(t)$ with log-spectra X_1 and X_2 respectively, and with $x_s(t) = x_1(t) + x_2(t)$ with log-spectrum $X_s \approx M \cdot X_1 + (1 - M) \cdot X_2$ where M comes from the element-wise maximum-indicator operator applied to the individual log spectrograms, $M = maxind(X_1, X_2)$, where

$$maxind(a,b) = \begin{cases} 1 & \text{when } a > b \\ 0 & \text{otherwise} \end{cases} \qquad (9)$$

Refiltering recovery consists of estimating a mask M_{est} from the composed log-spectrogram, recovering the individual sources from a composed audio signal by assigning a weight to each time-frequency bin of the composed signal spectrogram, i.e. $\hat{X}_1 = M_{est} \cdot X_s$ and $\hat{X}_2 = (1 - M_{est}) \cdot X_s$.

In [1], 8000 states HMMs were built for two different speakers. To analyze an unseen composed signal, the speaker models were composed into a factorial HMM (fig. 2b), in which composed observations are formed from a combination of the individual states of each HMM. The emission probabilities are given by:

$$p(x_t = y \mid s_t^1 = i, s_t^2 = j) = \mathcal{N}(x, max(\mu_i^1, \mu_j^2), C) \qquad (10)$$

where s_t^1 is the state of HMM 1, s_t^2 is the state of HMM 2, x_t is the composed observation, and $max(\cdot, \cdot)$ represents an element-wise-maximum operator.

Ideally, refiltering would be done by finding the factorial Viterbi path for the composed signals, which consists at each frame in the pair of states (one for each chain) that maximize the likelihood of the entire composed sequence. Given the pair of states $s_t^1 = i, s_t^2 = j$ from the factorial Viterbi path at time frame t, the mask m_t is found by applying the bitwise maximum operator to the means of the Viterbi states, $m_t = max(\mu_i, \mu_j)$. In practice, the true Viterbi path cannot be calculated due to the combinatorial explosion in the size of the factorial state space $N^2 = 8000^2 = 6.4 \times 10^7$. In [1], a limited set of factorial states with the highest observation likelihood at each time frame are used to perform Viterbi decoding on a limited grid. This approach does not guarantee that the solution found has the highest likelihood since it has a strong bias towards the observation probability.

We built multiband models for two speakers and combined them into a factorial model to explain new composed signals. The procedure is similar to the one discussed above, but using a full set of factorial emission probabilities (10) since in each band our state space is considerable smaller than when using a single 'full spectrum' HMM. This makes the combinatorial problem less daunting, and variational inference in the complete factorial state space can be performed. Refiltering is done by estimating the mask for band k as the maximum-indicator between the 'expectations' of the state means for each chain under the variational parameters taken as posteriors, i.e.:

$$M_t^k = maxind\left(\sum_j Q(S_t^{1k} = j) \cdot \mu_i^{1k}, \sum_j Q(S_t^{2k} = j) \cdot \mu_i^{2k}\right) \qquad (11)$$

The variational parameters are obtained in a iterative process that may involve a few passes over the entire multiband model.

5. EXPERIMENTS AND RESULTS

Models for two books-on-tape speakers were trained using 150,000 frames of speech, or about 40 minutes each. We built full-spectral HMMs with 1000 states to compare with multiband models with varying numbers of bands and coefficients per band. Subband models train much faster: Three EM iterations of the full-spectral HMMs for each speaker took over two weeks using the HTK software tools, whereas 20 iterations of the multiband model took in average 3 to 4 days using Matlab. The speaker models were tested in a refiltering source-separation task, where test samples of the two speakers were added together, and state sequences for each speaker were estimated via factorial HMM inference.

We quantify the degree of separation obtained by a given estimated mask, M_{est}, by measuring the Signal-to-Noise Ratio (SNR) of the resultant "separated" signals. The SNR for a given speaker measures the ratio of the content of the desired speaker versus the other speaker on the desired speaker "separated" output. When test

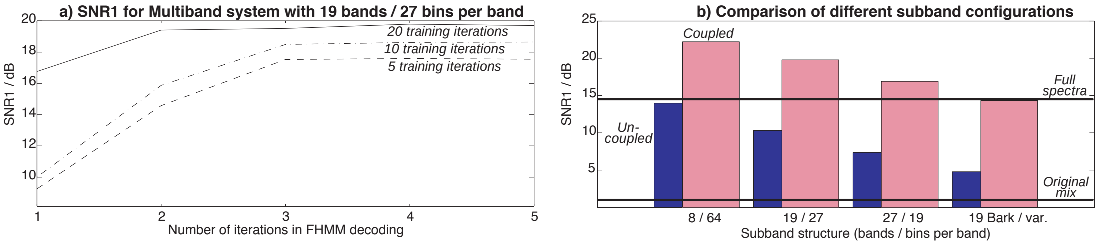

Fig. 3. a) SNR1 for a 19-band system versus iterations in recognition and training. b) SNR1 for different structures of independent and coupled multiband systems, where the first bar in each pair corresponds to the independent model and the second to the proposed coupled model.

signals are constructed by artificially summing individual source signals, it is possible to identify the portions of the final filtered mixture that properly originate in each source by passing the individual sources through the same time-varying filter. Using these separately-filtered components, the signal-to-noise (SNR) ratio is obtained by treating each reconstruction as a corrupt version of the original target signal, i.e. for speaker 1:

$$SNR_1 = 10 \cdot log_{10} \frac{\sum_{t,k} |X_1^*|^2}{\sum_{t,k} M_{est} \cdot |X_2^*|^2 + (1 - M_{est}) \cdot |X_1^*|^2} \qquad (12)$$

The noise denominator is obtained by direct subtraction of the original source from the refiltered output. This penalizes both inclusion of energy from the interference as well as deletion of target energy. X_1^* can be either X_1 (the log-spectra of $x_1(t)$), or $M_{opt} \cdot X_s$, the log-spectra obtained from the optimal mask $M_{opt} = maxind(X_1, X_2)$. We have observed that SNRs computed with the latter have a higher correlation with the perceptual quality of the separated signals. (Since M_{opt} is the best mask we can achieve under the model, in the sense of giving the best SNR against the original target, it also measures how close a given solution is to the best possible solution.)

Fig. 3a depicts SNR_1 values for 19 bands with 27 coefficients per band and 30 states per band, the vertical axis corresponds to the obtained SNR in dB, while the horizontal axis corresponds to the number of iterations performed on the factorial multiband model. There are three traces, corresponding to the SNR values obtained using speaker models trained for 5 (dashed line), 10 (dotted-dashed line), or 20 (solid line) iterations. We obtained a higher SNR when using parameters trained with more iterations, showing the benefits of the coupling. Fig. 3b shows four pairs of bars, each pair corresponds to a multiband model with a different structure (8 bands/64 coefficients, 19/27, 27/19 and 19 Bark-spaced bands with between 6 and 128 bins). The first bar in each couple corresponds to the SNR obtained when the bands are trained independently, the second band corresponds to the SNR obtained by the proposed model. The higher horizontal black line (around 14 dB) corresponds to the SNR obtained by the 1000-state full-spectrum model with a 100-state limited Viterbi grid. The lower line (1.2 dB) show the SNR calculated for the original mixture. We note that training the coupled models gives a consistent SNR improvement of around 10 dB in all models, with fewer, larger subbands (e.g. 8 bands of 64 bins) performing best. The Bark-scaled bands, which more closely reflect the perceptual information density of speech, are disappointing, but this may indicate the limited perceptual relevance of our SNR performance metric.

6. SUMMARY AND CONCLUSION

We have presented a new grid-like multiband model for acoustic modeling that enforces consistency between bands by coupling adjacent bands. The multiband models are capable of modeling rich and highly variable acoustic signals such as a person's speech with a relatively small number of spectral parameters. At the same time, the coupling between subbands preserves accuracy and consistency in the resulting complete spectral representations. The multiband model not only outperforms its full-spectral counterpart on the degree of separation achieved, but is also several orders of magnitude faster. We obtained interesting results for different structures of the multiband model.

7. ACKNOWLEDGMENTS

This work was supported by Microsoft Research and under NSF grant IIS-0238301, "The Listening Machine - Sound Source Organization for Multimedia Understanding".

8. REFERENCES

[1] S. Roweis, "One-microphone source separation," in *Advances in NIPS*, pp. 609–616. MIT Press, Cambridge MA, 2000.

[2] N. Mirghafori, *A Multiband Approach to Atomatic Speech Recognition*, Ph.D. thesis, Dept. of EECS, UC Berkeley, 1998.

[3] H. Bourlard and S. Dupont, "Subband-based speech recognition," in *Proc. ICASSP*, 1997.

[4] M. I. Jordan and C. Bishop, "Introduction to graphical models," 2002, In progress.

[5] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "An introduction to variational methods," in *Learning in Graphical Models*, M. I. Jordan, Ed. Kluwer Academic, 1997.

[6] R. Neal and G. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, M. I. Jordan, Ed. 1998, Kluwer.

[7] M. Reyes-Gomez, D. Ellis, and N. Jojic, "Subband audio modeling for single-channel acoustic source separation," Tech. Rep., Columbia/Microsoft, October 2003, `www.ee.columbia.edu/~mjr59/Multiband-TR.pdf`.

[8] S. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. EuroSpeech*, Geneva, 2003.