A GRAPHICAL MODEL FOR FORMANT TRACKING

Jonathan Malkin, Xiao Li, Jeff Bilmes

SSLI Lab, Department. of Electrical Engineering
University of Washington, Seattle

ABSTRACT

We present a novel approach to estimating the first two formants (F1 and F2) of a speech signal using graphical models. Using a graph that takes advantage of less commonly used features of Bayesian networks, both v-structures and soft evidence, the model presented here shows that it can learn to perform reasonably without large amounts of training data, even with minimal processing on the initial signal. It far outperforms a factorial HMM using the same assumptions and suggests that with further refinement the model may produce high quality formant tracks.

1. INTRODUCTION

Formants, the resonant frequencies of the vocal track during voiced speech, are widely believed to be useful features for both automatic speech recognition and speech synthesis [1]. Additionally, projects such as the UW Vocal Joystick, a new research effort at the University of Washington, are exploring the use of formants for 1-D and 2-D continuous motion control. The Vocal Joystick project is creating a new vocal interface to allow people, especially individuals with motor impairments, to use many aspects of their voice to easily interact with computers or other devices.

Formant tracking is a difficult problem for which there have been many proposed solutions. Many, such as [2], use LPC spectral analysis to estimate potential formant frequencies. As frame-based estimates relying on LPC tend to be noisy, post-processing is typically applied, often either continuity constraints through dynamic programming (DP) or template matching. There have also been other types of formant trackers such as HMM-based methods [3], approaches using nonlinear predictors [4], and a recent one using a Kalman filtering framework [5], to name a few. It should be noted that the last two actually aim to model the vocal track resonant frequencies more generally, that is during both voiced and unvoiced speech segments.

This paper presents a novel graphical model for use with formant tracking, inspired by the model in [6]. The goal of this project is to use data to learn a formant tracker, including both candidate estimation and continuity constraints, and to do so using a relatively simple model. Additionally, because of the nature of the graphical model framework, the method will be able to learn parameters for any set of provided features and with any additional constraints provided. This allows it to quickly adjust to customized tasks for which typical assumptions no longer hold; for instance, the Vocal Joystick project will allow formants to change arbitrarily, possibly violating typical assumptions about the rate of change in natural speech. Specifically, post processing based on dynamic programming, a widely used technique, relies on evaluating both a local cost function and a transition cost function [7]. Finding appropriate functions to use becomes a crucial part of successfully applying DP, but there can be many parameters to tune. By employing statistical models, we can learn the best values of those parameters, in a maximum likelihood sense, for any task.

This work was supported by NSF grant ITS-0326382.

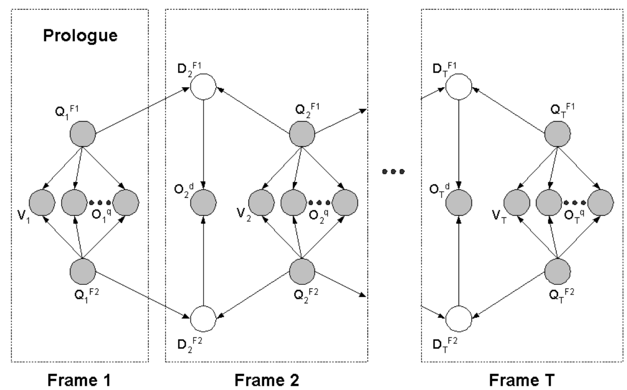

Fig. 1. V-structure sequence model training graph

While graphical models have been used before in formant tracking, there are several differences between those approaches and the one presented here. The HMM-based method of [3] uses a graphical model only to decide from among a set of pre-selected candidates and to learn continuity constraints. The Kalman filtering-based method of [5] is complex but with quite good results. By contrast, the graphs presented here exploit several useful properties of Bayesian networks, namely v-structures and the use of virtual or soft evidence [8, 9], unused by most other approaches. In doing so, we present a clear and elegant way of simplifying the statistical model.

2. GRAPHICAL MODEL STRUCTURE

2.1. The Training Graph

The power of graphical models comes from their ability to model families of probability distributions in a quick, concise, and easily understood manner. This paper uses a new model for training as shown in Figure 1, and ultimately compares the results of this graph with those of a factorial HMM (FHMM) [10]. Shaded nodes represent nodes observed during training and unshaded nodes represent hidden variables.

The most obvious difference between this graph and its decoding counterpart presented in the next subsection and a factorial HMM is its use of v-structures instead of directly linking consecutive states. Typical HMMs have no v-structures [8]; factorial HMMs do include them by using separate state nodes in the same frame as parents of observations. Our new graphs, in addition to the v-structures formed by the state and observation nodes, also include them between states in adjacent frames. In this way, the graphs approximate factorization possibilities in an undirected graphical model while remaining a Bayesian network.

The random variables Q_t^{F1} and Q_t^{F2} are the formant frequency states for $F1$ and $F2$ and time frame t. Each variable has cardinality N which represents $N-1$ possible frequencies plus an unvoiced state which is given index N. Neither variable has parents, but each has its own prior distribution: π_q^{F1} and π_q^{F2}.

Every time frame except frame 1 has a pair of hidden random variables, D_t^{F1} and D_t^{F2} with cardinality K, corresponding to the difference between adjacent formant states. Although there are a total of N^2 possible transitions between any pair of formants $(Q_{t-1}^{Fn}, Q_t^{Fn}), n \in \{1,2\}$, using these difference nodes quantizes those transitions to only K values; this provides only very rough quantization (we use Fn alone to refer to both $F1$ and $F2$). Additionally, this difference node is represented by a deterministic relationship given its parents, meaning it is calculated with a simple lookup table. The nodes for a given formant share the same table, but a separate table is used for each formant. Each table splits the $(N \times N)$-sized transition table into K disjoint subsets $S_k, k = 1..K$ where $P(D_t^{Fn} = k | Q_{t-1}^{Fn} = i, Q_t^{Fn} = j) = 1$ iff $(i,j) \in S_k$. Both tables use the mapping:

$$\begin{aligned} S_1 &= \{(i,j) : i = N, j \neq N\} \\ S_2 &= \{(i,j) : i \neq N, j = N\} \\ S_k &= \{(i,j) : m_k \leq j - i < M_k\};\ k = 3..K. \end{aligned} \tag{1}$$

S_1 and S_2 represent (resp.) unvoiced-to-voiced and voiced-to-unvoiced transitions. The quantities m_k and M_k are (resp.) the minimum and maximum index for the k^{th} subset, where each of the subsets are evenly spaced at integers between $-N+2$ and $N-2$ inclusive. These subsets form equivalence classes based on the difference between adjacent formant frequencies (so only formant transitions and not absolute frequency values are used). The unvoiced-to-unvoiced transition maps into the same subset as a voice transition with $i = j$. Note that the choice of parameters used in this mapping can constrain formant changes per frame, in contrast to allowing free variation as presented in the introduction.

This mapping disregards absolute formant values in favor of the relative change in value. In doing so, it compresses what would otherwise be a much larger table into one that is small enough to be learned even with limited amounts of training data. Also, when later applied to decoding, the information from the priors π_q^{F1} and π_q^{F2} along with observations will constrain the formant values sufficiently well that the transition table can be compressed with little or no detrimental impact.

The observed children of Q_t^{F1} and Q_t^{F2} are primarily continuous observation vectors O_t^q, with the exception of a binary voicing indicator V_t (an observation obtained using voicing estimates generated independent of this model; see Section 3.2 for specific details). There are also difference observations, O_t^d as children of both D_t^{F1} and D_t^{F2}. All features come from signal processing on the audio signal and will be discussed further in subsection 2.3.

The graph of Figure 1 should effectively capture the relationships between both the vocal tract resonances (in voiced speech) and the acoustics, as needed for formant tracking. Note that Q_t^{F1} and Q_t^{F2} are *not* independent due to the observations. This graph could be represented as a factorial HMM (or even a non-factorial HMM) where D_t^{Fn}, Q_{t-1}^{Fn} and Q_t^{Fn} are grouped into a single large node, but it would have very high cardinality and would consequently require much more data to train effectively.

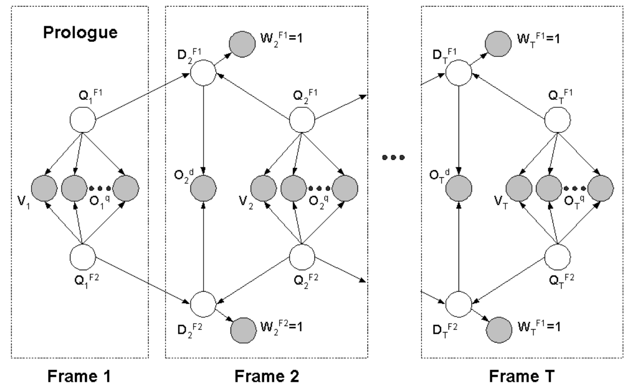

Fig. 2. V-structure sequence model decoding graph

2.2. The Decoding Graph

Figure 2 shows the graph used for decoding. It differs from the training graph in that the formant frequency states are now hidden. There is also an additional observed child W_t^{F1} for each D_t^{Fn} node with a constant value of 1. This allows the use of virtual or soft evidence for D_t^{Fn}, denoted by the conditional probability

$$\pi_d^{Fn}(k) \triangleq P(W_t^{Fn} = 1 | D_t^{Fn}), \tag{2}$$

which is in practice set equal to the normalized histogram of the K formant transition patterns. This allows us to capture a prior distribution over each difference node D_t^{Fn}, something which would be difficult to accomplish without the use of soft evidence. The normalization is not actually necessary for inference and decoding; π_d^{Fn} could be scaled by any nonnegative value and the results would be unchanged. Note that Q_{t-1}^{Fn} is *not* independent of Q_t^{Fn} due to the use of soft evidence.

2.3. Observation Features

As with many tasks, selecting appropriate observation features is crucial. In this case, variations on coefficients of the power spectrum computed from linear prediction coefficients were used as features for O_t^q directly (see also [6]). Two variations were tried: the logarithm of the power spectrum, and the logarithm of the power spectrum after it has been normalized to sum to 1.

Formants are likely to be at or near the peaks of the power spectrum for any time frame. Consequently, if a frequency value is near the peak, it is more likely to be a formant than one that is not. With this in mind, we define observation vector $O_t^q = (O_{1,t}^q, O_{2,t}^q, ..., O_{N-1,t}^q)$ where $O_{i,t}^q,\ i = 1..N-1$ is the power spectrum associated with the i^{th} frequency bin. Using this notation, we define an "improper" probability distribution for the observations as follows:

$$\begin{aligned} &P(O_t^q = x_t | Q_t^{F1} = i, Q_t^{F2} = j) \\ &= \begin{cases} 0, & i > j \\ 1, & i = j = N \\ \mathcal{N}(x_{i,t}; \mu_i, \gamma_1^2) \cdot \mathcal{N}(x_{j,t}; \mu_j, \gamma_2^2), & i,j = 1..N-1 \end{cases} \end{aligned} \tag{3}$$

while voicing is used to enforce constraints on the formants:

$$P(V_t | Q_t^{F1} = i, Q_t^{F2} = j) = \begin{cases} 1, & i \leq j < N, V_t \text{ voiced} \\ 0, & i,j \neq N, V_t \text{ unvoiced} \end{cases} \tag{4}$$

Note that some combinations are made impossible (given zero probability), for instance F1 being at a higher frequency than F2.

Also, via Eq. 4, when the signal is voiced, the model forces Q_t^{Fn} to take values in $\{1, ..., N-1\}$; when unvoiced the variables are forced to assume value N. In the unvoiced case, the score value is fixed at 1 (effectively turning off these observations). Thus, by forcing the model into state N during unvoiced regions, we only use Gaussian scores when tracking formants.

Different fixed means were used for the Gaussians depending on the observation features used. For the log of the power spectrum, the mean for the i^{th} frequency was set to the maximum value seen at that frequency in the training set, implying a different mean for each observation node. By contrast, for the log of the normalized power spectrum the mean was 0, the maximum possible value, for all frequencies. In both cases, a single covariance was trained and used for all Gaussians while holding these means fixed.

By fixing the mean of the Gaussians at the maximum value from the training data for each observation in the unnormalized case and the maximum possible value in the normalized case, we use only the lower half of the distribution. Due to the Gaussian's monotonicity in its lower half, frequencies with larger feature values are considered more likely to be formants than those with lower values. In the unnormalized case, a one-sided distribution would be preferable, otherwise we cannot ensure that a larger feature value is indeed treated as more likely. Using a Gaussian, though, we can calculate a variance to determine how fast the function should fall off. Since the maximum possible value in the unnormalized case is not known, even allowing the use of another improper distribution, a reasonable and trainable method for determining the shape of a one-sided distribution is not obvious.

Finally, in this initial work, we gave O_t^d probability 1 for all parent values, effectively removing it from the graph. This had the benefit of leading to a better decoding algorithm by allowing for an efficient triangulation.

3. EXPERIMENTAL FRAMEWORK AND RESULTS

3.1. Structure

The training and testing sets for this formant tracker were derived from two databases. The first is "Mocha-TIMIT," developed at Queen Margaret University College, and the other was originally created at the Hong Kong University of Science and Technology for tone-estimation research. Both databases consist of read English speech and include laryngograph data.

For each waveform, Entropic's *get_f0* [11] was used to extract pitch and voicing information from the laryngograph waveforms. Similarly, *formant* was used to obtain formant frequencies and bandwidths. These values were then fed into an updated version of the Klatt synthesizer [12], a formant-based speech synthesizer. In doing so, we obtained accurate formant labels despite lacking a hand-labeled corpus.

The training set comprised 1200 utterances, 300 from each of two male and two female speakers. The testing set used the same four speakers and amounted to 639 utterances. Equal numbers of utterances were used from each speaker in both sets.

The Graphical Model Toolkit [13] (GMTK), was used to implement the models for these experiments. The proposed model was compared to a FHMM, which used a 1st-order Markov chain between formant states rather than the v-structures of the proposed model. The FHMM used an identical improper distribution for the observation features.

3.2. Implementation Details

Observation features were generated using a custom front end. The input waveforms were sampled at 16kHz, from which 40ms frames were created, spaced 10ms apart. Light center clipping was applied to remove background noise. A 10th order LPC filter was used and the power spectrum calculated at 128 frequency points in the range 0Hz-4000Hz, a little wider than the range in which the first two formants exist. Those values became the observation features O_t^q. Adding an unvoiced state gave each node Q_t^{Fn} a cardinality of 129, and a frequency resolution of about 31Hz. The frequency transitions were quantized into 129 bins, although fewer could have been used with equal effectiveness or a more complex mapping used, as most bins were empty.

Formant	F1				
Features	log PSD		log norm. PSD		ESPS
Model	FHMM	New	FHMM	New	
GER	87.38%	78.32%	55.30%	50.33%	21.99%
Mean	71.72	-52.23	-116.36	-59.88	3.08
L1-norm	292.97	142.68	124.55	103.25	74.99

Formant	F2				
Features	log PSD		log norm. PSD		ESPS
Model	FHMM	New	FHMM	New	
GER	88.67%	31.12%	96.14%	30.94%	32.74%
Mean	362.93	-21.65	-19.45	123.20	285.90
L1-norm	1320.13	317.62	1228.36	317.06	424.70

Formant	Overall				
Features	log PSD		log norm. PSD		ESPS
Model	FHMM	New	FHMM	New	
GER	98.96%	84.57%	98.53%	64.83%	45.19%

Table 1. Gross error rates (GER, >20% difference) for F1, F2, and overall (frames with at least one error) for both sets of features. ESPS results are provided as a baseline. Also given where applicable are the mean error or bias in Hz and the zero-centered L1-norm in Hz.

Formant frequency priors, π_q^{F1} and π_q^{F2}, were smoothed because of the small size of the training set. By quantizing the transitions we are able to only lightly smooth the transition "priors" π_d^{F1} and π_d^{F2}. The FHMM used the same smoothed formant frequency priors as the new model. Its transition probabilities suffered from a lack of data, so we also used *SRILM* [14] to train a standard ARPA backoff bi-gram "language model" as the conditional probability table within GMTK.

In this paper, the goal was to train a formant tracker not to train a model for making a voicing decision. Consequently, oracle voicing information was used in both training and testing so as to avoid errors from an incorrect voicing decision and more effectively evaluate this model's formant tracking ability. This model could, of course, use the results of any voicing decision model.

3.3. Results

Results of running this model with each set of features appear in Table 3.2, with results from ESPS's *formant* included as a baseline.

The best results for the new model are nearly 2% better than ESPS's on an absolute scale for F2 tracking, a significant improvement. The new model always outperforms the factorial HMM when measuring error rate. In one case the FHMM has a lower mean, but its norm is vastly larger; this simply means the many large errors are more evenly distributed between being too high and too low. It should be noted that the FHMM results are for the unsmoothed transition probabilities as the bi-gram model gave slightly worse results which are not presented here. The new model also outperforms ESPS on F2 with both sets of features, although by a smaller margin with the unnormalized features. The normalized features are better for the new model, and slightly better overall for the FHMM even though it had a worse F2 result using them.

Also, the new model's average error magnitude, measured via

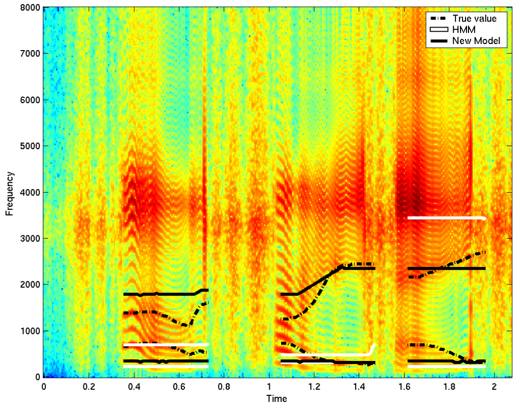

Fig. 3. Spectrogram with overlaid formants

the zero-centered L1-norm, $\sum_{voiced\,frames} \frac{|Fn_{true} - Fn_{decoded}|}{count(voiced\,frames)}$, is substantially smaller than those of the FHMM in all cases. When it beats ESPS, the new model also has lower bias and average error magnitude values; that is not the case when ESPS performs better.

ESPS's result is superior to that of either model overall, largely because of its much superior F1 results. There are several possible reasons for this which will be addressed in the final section.

Figure 3 shows a spectrogram of one of the synthesized speech signals with formants overlaid. F2 is not typically right on, but often appears to be moving in the right direction compared to the formant. The learned transition probabilities seem too constraining as it generally moves more slowly than the actual formant and occasionally decides that not moving is best. The FHMM, however, is way off the mark. For F1, both the HMM and new model appear to be confused by the fundamental frequency and to decide that deviating from there is too expensive.

We also ran preliminary experiments lower-bounding F1 by pitch (oracle information). Only F1 was affected by this change. There was a small improvement in error rate for the unnormalized features, and no change for the normalized features. A closer look at the results revealed that for the unnormalized features, predicted F1 dropped below pitch at some point in two-thirds of utterances by a female speaker and never for male speakers. By contrast, using the normalized features this happened only on about one-quarter of the utterances by female speakers, and for a shorter average duration than when using unnormalized features.

4. DISCUSSION

The results for the new model are quite encouraging. Minimal signal processing was performed on the input speech waveforms before calculating power spectra, and the model was able to learn reasonable F2 behavior. Tracking F1 is still something of a challenge, though. One reason, as is more apparent from the spectrogram, is that the models, both the new one and the FHMM, are confusing the fundamental frequency with F1. Using the spectrum from an LPC filter means that that pitch's effects may bleed into adjacent frequencies, so simply constraining F1 by pitch is not a guaranteed solution.

Another potential problem is that F1 tends to vary around a smaller range than does F2. By calculating the power spectrum at only 128 points, the resolution in the lower parts of the spectrum may be insufficient for tracking the first formant. Increasing the resolution is an option, but it would increase the cardinality of the graph which could be problematic. Instead, we could use a logarithmic frequency scale, concentrating more resolution at lower frequencies. This may have detrimental side-effects on tracking F2, though. We are currently experimenting with this variation.

Additionally, many formant trackers use other knowledge of formants to help select candidates whereas this system uses only the frequency. Specifically, formant bandwidths are commonly used to screen out potential candidates since true formants are fairly concentrated in frequency. To extend this further, it may be necessary to pre-select some set of possible candidates and then use the graphical model to evaluate those options.

Despite these challenges, to be addressed by future work, the new model quite clearly holds much promise. Its strongly superior performance versus the factorial HMM demonstrates that it is able to learn a useful model from a very limited amount of training data and suggests it will play a useful role in the Vocal Joystick project..

The authors would like to thank Chris Bartels and Gang Ji for much help with GMTK.

5. REFERENCES

[1] X.Huang, A.Acero, and H.-W.Hon, *Spoken Language Processing*, Prentice Hall PTR, 2001.

[2] K.Xia and C.Espy-Wilson, "A new strategy of formant tracking based on dynamic programming," in *Proc. Int. Conf. on Spoken Language Processing*, 2000.

[3] A.Acero, "Formant analysis and synthesis using hidden markov models," in *Proc. Eur. Conf. Speech Communication Technology*, 1999.

[4] L.Deng, I.Bazzi, and A.Acero, "Tracking vocal tract resonances using an analytical nonlinear predictor and a target-guided temporal constraint," in *Proc. Eur. Conf. Speech Communication Technology*, 2003.

[5] L.Deng, L.Lee, H.Attias, and A. Acero, "A structured speech model with continuous hidden dynamics and prediction-residual training for tracking vocal track resonances," in *IEEE ICASSP*, 2004.

[6] X.Li, J.Malkin, and J.Bilmes, "Graphical model approach to pitch tracking," in *Proc. Int. Conf. on Spoken Language Processing*, 2004.

[7] H.Ney, "Dynamic programming algorithm for optimal estimation of speech parameter contours," *IEEE Trans. on Systems, Man and Cybernetics*, vol. SMC-13, no. 3, pp. 208–214, 1983.

[8] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 2nd printing edition, 1988.

[9] J.Bilmes, "On soft evidence in bayesian networks," Tech. Rep. UWEETR-2004-0016, U. Washington Dept. of Electrical Engineering, 2004.

[10] Z.Ghahramani and M.Jordan, "Factorial hidden markov models," *Machine Learning*, vol. 29, pp. 245–275, 1997.

[11] D.Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W.B.Kleign and K.K.Paliwal, Eds., Amsterdam, 1995, pp. 495–515, Elsevier Science.

[12] D.Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Soceity of America*, vol. 67, pp. 971–995, 1980.

[13] J.Bilmes and G.Zweig, "The Graphical Models Toolkit: An open source software system for speech and time-series processing," in *IEEE ICASSP*, 2002.

[14] A.Stolcke, "SRILM – an extensible language modeling toolkit," in *Proc. Int. Conf. on Spoken Language Processing*, 2002, vol. 2, pp. 901–904.