Learning Prosodic Sequences Using the Fundamental Frequency Variation Spectrum

Kornel Laskowski[1]*, Jens Edlund*[2]*, and Mattias Heldner*[2]

[1] interACT, Carnegie Mellon University, Pittsburgh PA, USA
[2] Centre for Speech Technology, KTH, Stockholm, Sweden

kornel@cs.cmu.edu

Abstract

We investigate a recently introduced vector-valued representation of fundamental frequency variation, whose properties appear to be well-suited for statistical sequence modeling. We show what the representation looks like, and apply hidden Markov models to learn prosodic sequences characteristic of higher-level turn-taking phenomena. Our analysis shows that the models learn exactly those characteristics which have been reported for the phenomena in the literature. Further refinements to the representation lead to a 12-17% relative improvement in speaker change prediction for conversational spoken dialogue systems.

1. Introduction

While speech recognition systems have long ago transitioned from formant localization to spectral (vector-valued) formant representations, prosodic processing continues to rely squarely on a pitch tracker's ability to identify a single peak, corresponding to the fundamental frequency. Unfortunately, peak localization in acoustic signals is particularly prone to error, and pitch trackers (cf. [9]) and downstream speech processing applications [10] employ dynamic programming, non-linear filtering, and linearization to improve robustness. These long-term constraints violate the temporal locality of the estimate, whose measurement error may be better handled by statistical modeling than by (linear) rule-based schemes. But even if a robust, local, continuous, statistical estimate of absolute pitch were available, applications require instead a representation of prosody, or pitch *variation*, and they go to considerable additional effort to identify a speaker-dependent quantity for normalization.

In the current work, we revisit a recently-derived representation of fundamental frequency variation [8], which implicitly addresses most if not all of the above issues. Its properties make it particularly suitable to bottom-up, continuous statistical sequence learning. We evaluate the representation using an improved filterbank design, but most importantly we explore, for the first time, what the representation looks like, visually, and what statistical sequence models learn when presented with a specific higher-level target. Here, as in previous work [3][8], that target is the prediction of speaker change in the context of a conversational spoken dialogue system with a short (0.3s) response time [1][4][2].

2. Human-Human Dialogue Corpus

In an effort to endow conversational dialogue systems with human-like responsiveness, we study dialogues from the Swedish Map Task Corpus [7], which differ significantly from less interactive domains such as ATIS (cf. [6]). The data, shown in Table 1, has been divided into a DEVSET and an EVALSET which are disjoint in speakers.

Data Set	Duration (mn:ss)	Dialogue role g		
		speakers	# EOTs	# SCs
DEVSET	77:40	F4,F5,M2,M3	480	222
EVALSET	60:39	F1,F2,F3,M1	317	149

Table 1: Size, speakers (F=female, M=male), number of end-of-talkspurt (EOT) and speaker change (SC) events for the speaker in role g in our datasets.

Here, as in our previous work [3][8], we use the presence of an *observed* speaker change as the gold standard. Vocalization by speaker g is marshalled into talkspurts, separated by pauses at least 0.3s long, as hypothesized by an automatic speech activity detection component. At each end-of-talkspurt (EOT) event at time t, we investigate the behavior of g and of interlocutor f. If g's next talkspurt begins at $t + T^t_{g,\mathcal{N}}$, and f's next talkspurt begins at $t + T^t_{f,\mathcal{N}}$, we assign to the EOT the label

$$L_t \quad = \quad \begin{cases} \text{SC} & \text{if } T^t_{f,\mathcal{N}} - T^t_{g,\mathcal{N}} < 0 \\ \neg\text{SC}, & \text{otherwise} \end{cases} \qquad (1)$$

where SC is a *speaker change* and $\neg SC$ is *not a speaker change*. Online estimation of appropriateness to vocalize at time t by the system (attempting to mimic f's behavior) consists of predicting the value of L_t given only a prosodic description of $\mathbf{x}$, the last 500 ms of g's speech terminating at time t.

3. Fundamental Frequency Variation

In [8], we derived a spectral representation of near-instantaneous variation in fundamental frequency, which we will refer to here as the fundamental frequency variation (F0V) spectrum. Every 8 ms, we compute an $N{\equiv}512$-point Fourier transform over the left and right halves of a 32 ms frame, leading to frequency representations $\mathbf{F}_L$ and $\mathbf{F}_R$, respectively. The peaks of the two windows are 8 ms apart. The F0V spectrum is then given by

$$g^{\rho}[r] = \begin{cases} \dfrac{\sum |\tilde{F}_L\left(2^{-4r/N}k\right)| \cdot |F^*_R[k]|}{\sqrt{\sum |\tilde{F}_L\left(2^{-4r/N}k\right)|^2 \cdot \sum |F^*_R[k]|^2}}, & r \geq 0 \\ \dfrac{\sum |F_L[k]| \cdot |\tilde{F}^*_R\left(2^{+4r/N}k\right)|}{\sqrt{\sum |F_L[k]|^2 \cdot \sum |\tilde{F}^*_R\left(2^{+4r/N}k\right)|^2}}, & r < 0 \end{cases} \qquad (2)$$

where, in each case, summation is from $k = -N/2 + 1$ to $k = N/2$; for convenience, r varies over the same range as k.

The interpolated values $\tilde{F}_L$ and $\tilde{F}_R$ are given by

$$\begin{aligned}\tilde{F}_L\left(2^{-\rho}k\right) &= \left|\lceil 2^{-\rho}k\rceil - 2^{-\rho}k\right| F_L\left[\lfloor 2^{-\rho}k\rfloor\right] \quad (3)\\ &\quad + \left(1-\left|\lceil 2^{-\rho}k\rceil - 2^{-\rho}k\right|\right) F_L\left[\lceil 2^{-\rho}k\rceil\right] \ ,\\ \tilde{F}_R\left(2^{+\rho}k\right) &= \left|\lceil 2^{+\rho}k\rceil - 2^{+\rho}k\right| F_R\left[\lfloor 2^{+\rho}k\rfloor\right] \quad (4)\\ &\quad + \left(1-\left|\lceil 2^{+\rho}k\rceil - 2^{+\rho}k\right|\right) F_R\left[\lceil 2^{+\rho}k\rceil\right] \ ,\end{aligned}$$

and normalization in Equation 2 ensures that $g^\rho[r]$ is an energy-independent representation.

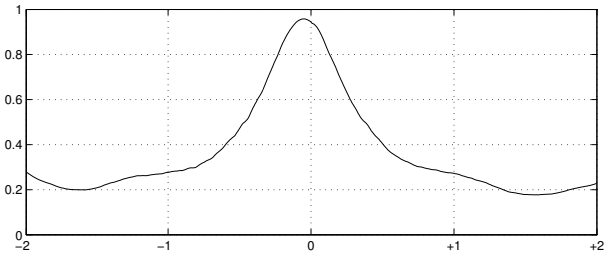


Figure 1: A sample fundamental frequency variation spectrum; the $x-$axis represents change in octaves per 8 ms.

A sample F0V spectrum, for a voiced frame, is shown in Figure 1; for unvoiced frames, the peak tends to be much lower and the tails much higher. The position of the peak, with respect to $r = 0$, indicates the current rate of fundamental frequency variation. However, rather than locating this peak, we utilize the representation as is, following the application of a filterbank, in subsequent modeling. We show, in the middle panel of Figure 2, what a "spectrogram" represenation looks like when F0V spectra from consecutive frames are stacked alongside one another.

4. Baseline Detector

The fundamental frequency variation spectrum $g^\rho[r]$ is computed every 0.008s over the 500ms preceding each EOT. Feature extraction for EOT classification (as either a SC or a $\neg SC$) consists of passing the fundamental frequency variation spectrum through a filterbank FBOLD, of which three filters are shown in Figure 3. The filterbank also contains two rectangular "extremity" filters with spans of $(-2,-1)$ and $(+1,+2)$ octaves per 0.008 seconds, as explained in [8]. This leads to a compressed representation of 5 scalars per frame.

For training models, the input space of 5 scalars per frame is centered and variance-normalized (VN) over the training set to have unity variance in each dimension. Alternately, we have applied whitening of the training data (following centering) via the Karhunen-Loéwe transform (KLT). For both the VN and KLT normalizations, the transforms (and the mean) are computed using the training data only, and the fixed training set transform, and subtraction of the fixed training set mean, are applied during test set classification.

Using the normalized representation, we train one fully-connected hidden Markov model (HMM) for each of the two classes, with each model $\mathcal{M}$ consisting of 4 states and one Gaussian per state; the Gaussian centers are initialized using K-means. Classification is then performed using

$$L_t = \arg\max_k P\left(\mathbf{x}\,|\,\mathcal{M}_k\right) \quad (5)$$

where k is either SC or $\neg SC$.

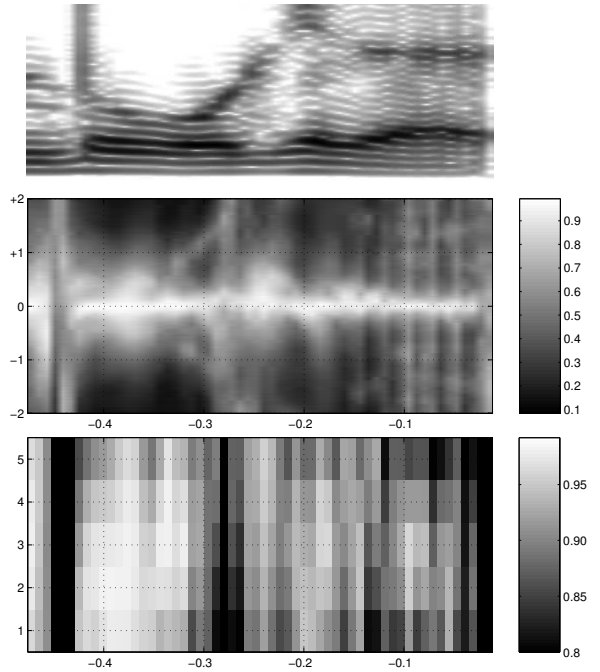


Figure 2: Spectrogram for a 500ms fragment of pre-EOT audio (top panel, upper frequency of 2kHz); F0V spectrogram (middle panel) for same fragment; and F0V spectrogram passed through the FBNEW filterbank shown in Figure 3.

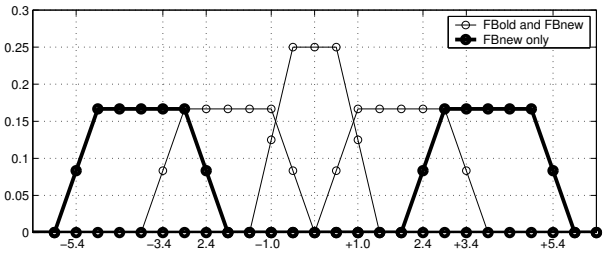


Figure 3: Three filters in the baseline filterbank (FBOLD) and in the improved filterbank (FBNEW), and the two additional filters used in FBNEW. The x-axis is in octaves per second; note that the filterbank is applied to frames in which $\mathbf{F}_L$ and $\mathbf{F}_R$ are computed at instants separated by 8 ms. The two extremity filters are not shown.

To assess classifier sensitivity to initialization, we train 10 HMMs for each of the SC and $\neg SC$ classes. We have noted that when data is whitened (KLT), initialization via K-means results in more dissimilar models than when VN normalization is used. Under these conditions, the intersection of the 100 hyperplanes induced by all $1 \leq i \leq 10$ HMMs for SC and $\neg SC$,

$$L_t = \arg\max_k \prod_{i=1}^{10} P\left(\mathbf{x}\,|\,\mathcal{M}_{k,i}\right) \quad (6)$$

leads to superior performance, often in excess of that achieved using the single best-performing $SC/\neg SC$ model pair.

5. Improved Filterbank

We present an improvement over the baseline obtained by modifying the filterbank design. The results shown here are the outcome of numerous leave-one-speaker-out round-robin experiments on the DEVSET, involving various modifications to the filterbank structure. The most significant improvement was obtained by extending the number of filters by 2, as shown in Figure 3, to yield a 7-filter filterbank FBNEW.

System	w/ VN		w/ KLT	
	mean	prod	mean	prod
EOV FBOLD	5.8	6.6	7.8	8.9
EOV FBNEW	9.7	10.3	7.7	10.3
EOT FBOLD	5.3	5.7	4.6	5.5
EOT FBNEW	7.1	7.9	6.0	6.6

Table 2: Discrimination on the DEVSET; numbers represent the area between an ROC curve and random guessing (max. is 50). "mean" and "prod" represent classifiers as in Equations 5 and 6, respectively.

We present a summary of these experiments in Table 2. Numbers represent the discrimination of the classifier, ie. the area between the diagonal (equal false and true positive rates) and the classifier's receiver operating characteristic (ROC). We show both the mean area, computed over all 100 hyperplanes induced by the 10 SC HMMs and the 10 $\neg SC$ HMMs (using Equation 5), and the area when the product of the model likelihoods is used (Equation 6). We also evaluate both the baseline filterbank FBOLD and the new filterbank in two conditions: (1) the EOT condition, in which **x** is the 500ms of audio immediately preceding the EOT; and (2) the end-of-voicing (EOV) condition, in which **x** is the 500ms of audio immediately preceeding the last voiced frame before the EOT. Voicing is determined using the Snack Sound Toolkit[1]; the EOV condition is used for contrast only.

As Table 2 shows, the product classifiers (Equation 6) always outperform individual classifiers (Equation 5) on the development set, and the difference in performance is larger for KLT-normalized features than for VN-normalized features, as mentioned in the previous section. In almost all experiments, FBNEW yields significantly better performance than FBOLD; the exception is the "mean" classifier with KLT-normalized features in the EOV condition, for which performance is not significantly different from the baseline.

System	w/ VN		w/ KLT	
	mean	prod	mean	prod
EOV FBOLD	13.4	13.7	14.4	16.5
EOV FBNEW	14.7	14.6	13.6	18.4
EOT FBOLD	14.6	15.2	14.5	16.6
EOT FBNEW	11.9	13.1	17.0	19.5

Table 3: Discrimination on the EVALSET; abbreviations as in Table 2.

We show in Table 3 the same experiments, conducted by training classifiers on all of the DEVSET and applying them to each speaker in the EVALSET. We note similar trends as for the DEVSET; however, in the EOT condition for VN-normalized features, FBNEW does not perform as well as FBOLD. We also note that performance on the EVALSET set is much better than on the DEVSET; we attribute this to the amount of training material, which is 33% larger in the EVALSET case (cf. Table 1). There may also be gender dependencies in the F0V representation or in the use of prosody; these issues are the subject of our ongoing analysis.

[1] http://www.speech.kth.se/snack/

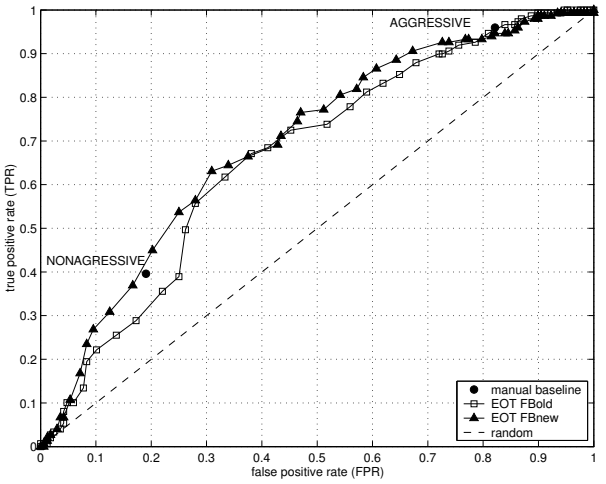


Figure 4: ROC curves for the FBOLD and FBNEW EOT systems.

In [8], we reported the EOV and EOT numbers for the evaluation set using only KLT-normalized features. With respect to this condition, FBNEW represents a relative performance improvement of 12% and 17%, respectively. We show the full ROC curve for the EOT system (with KLT-normalized features) in Figure 4, for both FBOLD and FBNEW, as well as for the hand-crafted baseline from [3]. As can be seen, FBNEW offers improved performance over FBOLD, over a significant range of possible true positive rates.

6. Model Analysis

Finally, we analyze what is actually learned by the HMM density estimators for the SC and $\neg SC$ classes. For simplicity, we look at the VN models; rotation via the Karhunen-Loéwe transform renders model-space features difficult to interpret. The topologies for a randomly chosen SC/$\neg SC$ model pair, with their transition probabilities, are shown in Figure 5.

We note first of all that the topologies, as learned for both classes, are quite similar. Sequences belonging to both classes appear to terminate in a visually identical state A, in which the proportion of energy in the harmonics is lower than for either states C or D in both topologies. Both SC and $\neg SC$ models contain a state B, with visually identical emission probabilities, which appears to capture unvoiced frames.

While the emission probabilities in states A, B, and D appear visually identical for both classes, the eccentricity of the locus of means of state C for the SC model appears slightly more pronounced that for the $\neg SC$ model. As this is hard to see in Figure 5, we show in Figure 6 the feature space, only for the C states, as it appears during training, *following* feature centering and variance normalization. It can be seen, by comparing diagrams (a) and (b), that the $\neg SC$ model is much

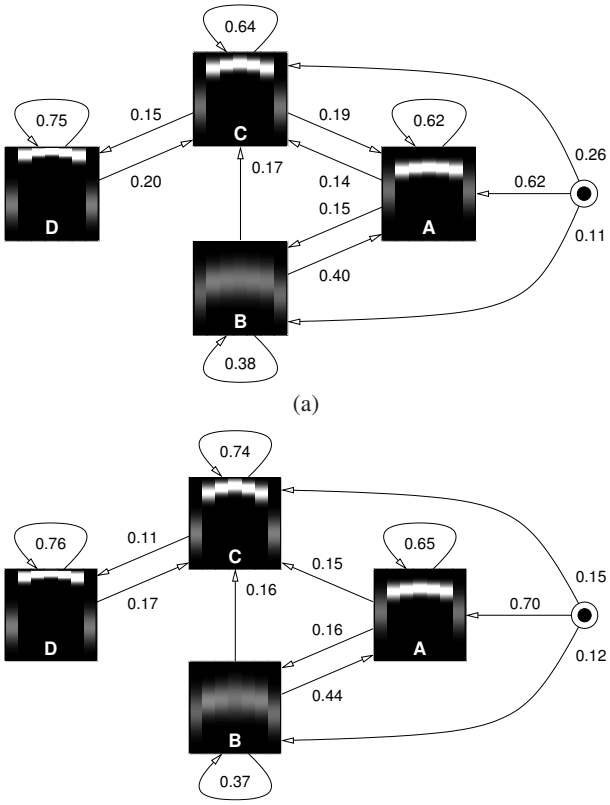
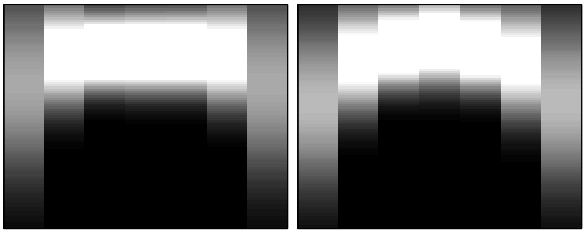


Figure 5: Transition and (unnormalized) emission probabilities as learned for (a) the SC model, and (b) the $\neg SC$ model; transition probabilities < 0.10 are not shown. Decoding begins with the most recent (latest) frame, and proceeds backwards in time.

more selective than the SC model about which filterbank output contains the maximum. This indicates that state C in the SC model accounts for flat, rising, and falling fundamental frequency contours, whereas the same state in the $\neg SC$ models clearly accounts for predominantly flat contours, in which the center filterbank output is largest in magnitude. This finding, that $\neg SC$ EOVs contain predominantly flat fundamental frequency contours while SC EOVs do not, corroborates numerous turn-taking studies, and was the main design principle behind the construction of our hand-crafted baseline [3].



(a) (b)

Figure 6: The distribution over normalized filterbank outputs learned for state C in (a) the SC model, and (b) the $\neg SC$ model.

7. Conclusions

Building on previous work [8] in which the fundamental frequency variation spectrum was derived, we have for the first time graphically demonstrated both the form of the representation, and its evolution in time. These, as well as our HMM classification results on an important speech-processing task, suggest that this representation is suitable for direct, principled, continuous sequence modeling such as that used in automatic speech recognition, not requiring peak-identification, dynamic programming, median filtering, landmark detection, linearization, or mean pitch estimation and subtraction. We have shown that, presented only with how humans behave, standard machine learning approaches using this representation allow an automated agent to learn to avoid locations to speak which are also avoided by humans, based on F0 variation alone. We have also demonstrated that the models which are actually learned corroborate research on human behavior. Finally, we have improved the filterbank design used in the compression of the fundamental frequency variation spectrum to yield relative performance improvements of 12-17% on held-out data.

8. Acknowledgments

We would like to thank Tanja Schultz and Rolf Carlson for continued encouragement of this collaboration. The work presented here was funded in part by the Swedish Research Council (VR) project 2006-2172, and in part by DARPA under contract HR001-06-2-001.

9. References

[1] Bell, L., Boye, J., and Gustafson, J, 2001. Real-time handling of fragmented utterances. In *Proc. NAACL 2001 Workshop on Adaptation in Dialogue Systems*, Pittsburgh PA, USA, 2-8.

[2] Skantze, G. and Edlund, J., 2004. Robust interpretation in the Higgins spoken dialogue system. In *Proc. ROBUST 2004*, Norwich, UK.

[3] Edlund, J. and Heldner, M., 2005. Exploring prosody in interaction control. In *Phonetica*, 62, 215-226.

[4] Ferrer, L., Shriberg, E., and Stolcke, A., 2002. Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody in human-computer dialog. In *Proc. ICSLP 2002*, Denver CO, USA, 2061-2064.

[6] Oviatt, S.L., 1996. Multimodal interfaces for dynamic interactive maps. In *Proc. CHI 1996*, New York NY, USA, 95-102.

[7] Helgason, P., 2006. SMTC - A Swedish Map Task corpus. In *Proc. Fonetik 2006*, Lund, Sweden, 57-60.

[8] Laskowski, K., Edlund, J. and Heldner, M., 2008. An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems. In *ICASSP 2008*, Las Vegas NV, USA.

[9] de Cheveigné, A. and Kawahara, H., 2002. YIN, a fundamental frequency estimator for speech and music. In *JASA*, 111:1917.

[10] Shriberg, E. and Stolcke, A., 2004. Direct modeling of prosody: An overview of applications in automatic speech processing. In *Proc. Speech Prosody 2004*.