First published in:
Proceedings, FONETIK 2008, Department of Linguistics, University of Gothenburg

The fundamental frequency variation spectrum

Kornel Laskowski[1], Mattias Heldner[2] and Jens Edlund[2]
[1]interACT, Carnegie Mellon University, Pittsburgh PA, USA
[2]Centre for Speech Technology, KTH Stockholm, Sweden

Abstract

This paper describes a recently introduced vector-valued representation of fundamental frequency variation – the FFV spectrum – which has a number of desirable properties. In particular, it is instantaneous, continuous, distributed, and well-suited to application of standard acoustic modeling techniques. We show what the representation looks like, and how it can be used to model prosodic sequences.

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 peak, corresponding to the fundamental frequency (F0) of the speaker. Peak localization in acoustic signals is particularly prone to error, and pitch trackers (cf. de Cheveigné & Kawahara, 2002) and downstream speech processing applications (Shriberg & Stolcke, 2004) employ dynamic programming, non-linear filtering, and linearization to improve robustness. These methods introduce long-term dependencies which violate the temporal locality of the F0 estimate, whose measurement error may be better handled by statistical modeling than by (linear) rule-based schemes. Even *if* a robust, local, analytic, statistical estimate of absolute pitch were available, applications require a representation of *pitch variation* and go to considerable additional effort to identify a speaker-dependent quantity for normalization (e.g. Edlund & Heldner, 2005).

In the current work, we describe a recently derived representation of fundamental frequency variation (see also Laskowski, Edlund, & Heldner, 2008a, 2008b; Laskowski, Wölfel, Heldner, & Edlund, in press), which implicitly addresses most if not all of the above issues. This spectral representation, which we will refer to here as the *fundamental frequency variation (FFV) spectrum* is (1) instantaneous, not relying on adjacent frames; (2) continuous, defined for all frames; (3) distributed; and (4) potentially sparse, making it suitable for the application of standard acoustic modeling techniques including bottom-up, continuous statistical sequence learning.

In previous work, we have shown that this representation is useful for modeling prosodic sequences for prediction of speaker change in the context of conversational spoken dialogue systems (Laskowski et al., 2008a, 2008b); however, the representation is potentially useful for any prosodic sequence modeling task.

The fundamental frequency variation spectrum

Instantaneous variation in pitch is normally computed by determining a single scalar, the fundamental frequency, at two temporally adjacent instants and forming their difference. F0 represents the frequency of the first harmonic in a spectral representation of a frame of audio, and is undefined for signals without harmonic structure. In the context of speech processing applications, we view the localization of the first harmonic and the subsequent differencing of two adjacent estimates as a case of suboptimal feature compression and premature inference, since the goal of such applications is not the accurate estimate of pitch. Instead, we want to leverage the fact that *all* harmonics are equally spaced in adjacent frames, and use *every* element of a spectral representation to yield a representation of the F0 delta.

To this end, we propose a vector-valued representation of pitch variation, inspired by *vanishing-point perspective,* a technique used in architectural drawing and grounded in projective geometry. While the standard inner product between two vectors can be viewed as the summation of pair-wise products with pairs selected by orthonormal projection onto a point at infinity, the proposed vanishing-point product induces a 1-point perspective projection onto a point at τ (Figure 1). When applied to two vectors representing a signal's spectral content, F_L and F_R, at two temporally adjacent instants, the vanishing-point product yields the standard dot product between F_L and a dilated version of F_R, or between F_R and a dilated version of F_L, for positive and negative values of τ, respectively.

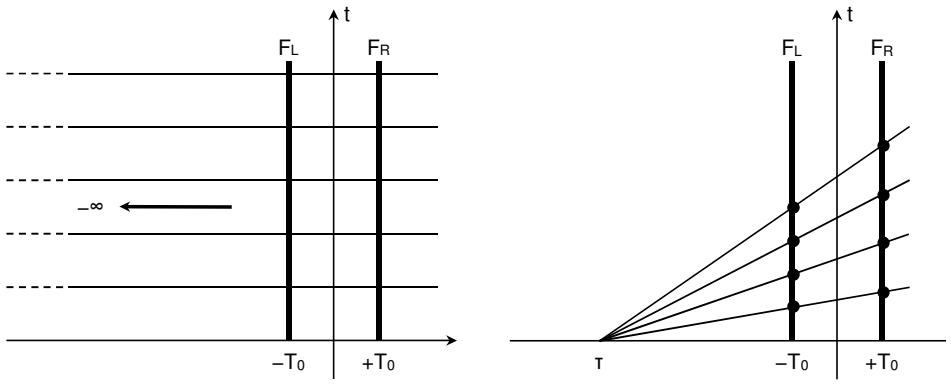

Figure 1. The standard dot-product shown as an orthonormal projection onto a point at infinity (left panel), and the proposed vanishing-point product, which generalizes to the former when $\tau \rightarrow \pm\infty$ (right panel).

The degree of dilation is controlled by the magnitude of τ. The proposed vector-valued representation of pitch variation is the vanishing-point product, evaluated over a continuum of τ. For each analysis window, centered at time t, we compute the short-time frequency representation of the left-half and the right-half portion of the window, leading to F_L and F_R, respectively, using two asymmetrical windows which are mirror images of each other, as shown in Figure 2.

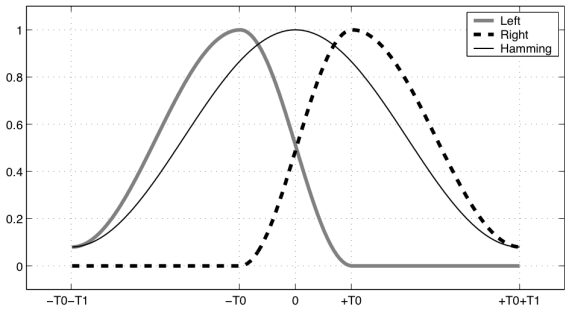

Figure 2. Left and right windows used for the computation of F_L and F_R, respectively, consisting of asymmetrical Hamming and Hann window halves. T_0 is 4 ms, and T_1 is 12 ms, for a full analysis window width of 32 ms. A 32 ms Hamming window is shown for comparison.

F_L and F_R are N≡512-point Fourier transforms, computed every 8. The peaks of the two windows are 8 ms apart. The FFV spectrum is then given by

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

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. Normalization ensures that *g[r]* is an energy-independent representation. The frequency-scaled, interpolated values $\tilde{F}_L$ and $\tilde{F}_R$ are given by

$$\tilde{F}_L(2^{-\rho}k)=\alpha_L F_L\left[\left\lfloor 2^{-\rho}k\right\rfloor\right]+(1-\alpha_L)F_L\left[\left\lceil 2^{-\rho}k\right\rceil\right],$$
$$\tilde{F}_R(2^{+\rho}k)=\alpha_R F_R\left[\left\lfloor 2^{+\rho}k\right\rfloor\right]+(1-\alpha_R)F_R\left[\left\lceil 2^{+\rho}k\right\rceil\right],$$

where

$$\alpha_L=\left|\left\lceil 2^{-\rho}k\right\rceil-2^{-\rho}k\right|,$$
$$\alpha_R=\left|\left\lceil 2^{+\rho}k\right\rceil-2^{+\rho}k\right|.$$

A sample FFV spectrum, for a voiced frame, is shown in Figure 3; 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. The sample FFV spectrum shown in Figure 3 thus indicates a single frame with a slightly negative slope, that is a slightly falling pitch.

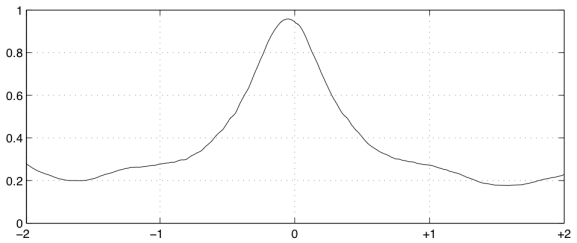

Figure 3. A sample fundamental frequency variation spectrum. The x-axis is in octaves per 8ms.

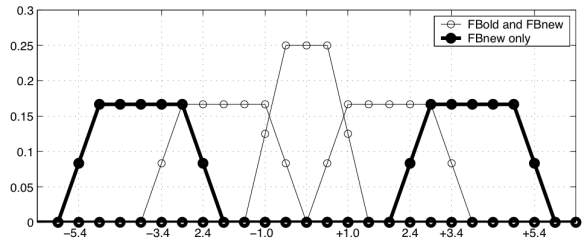

Figure 4: Filters in two versions of the filterbank. The x-axis is in octaves per second; note that the filterbank is applied to frames in which F_L and F_R are computed at instants separated by 0.008s. Two extremity filters at (−2, −1) and (+1, +2) octaves per frame are not shown.

Filterbank

Rather than locating the peak in the FFV spectrum, we utilize the representation as is, and apply a filterbank. The filterbank (FBNEW shown in Figure 4) attempts to capture meaningful prosodic variation, and contains a conservative trapezoidal filter for perceptually "flat" pitch ('t Hart, Collier, & Cohen, 1990); two trapezoidal filters for "slowly changing" pitch; and two trapezoidal filters for "rapidly changing" pitch. In addition, it contains two rectangular extremity filters with spans of (−2, −1) and (+1, +2) octaves per frame, as we have observed that unvoiced frames have flat rather than decaying tails. This filterbank reduces the input space to 7 scalars per frame.

We show what a "spectrogram" representation looks like when FFV spectra from consecutive frames are stacked alongside one another, in Figure 5, as well as what the representation looks like after being passed through filterbank FBNEW of Figure 4.

Modeling FFV spectra sequences

In order to transition from vectors of frame-by-frame FFV spectra passed through a filterbank to something more like what we normally associate with prosody, such as flat, falling, and rising pitch movements, *sequences* of FFV spectra need to be modeled. A standard option for modeling sequences involves training hidden Markov models (HMM). In previous work, we have used fully-connected hidden Markov models (HMM) consisting of four states with one Gaussian per state (see Figure 6). However, other HMM topologies are also possible.

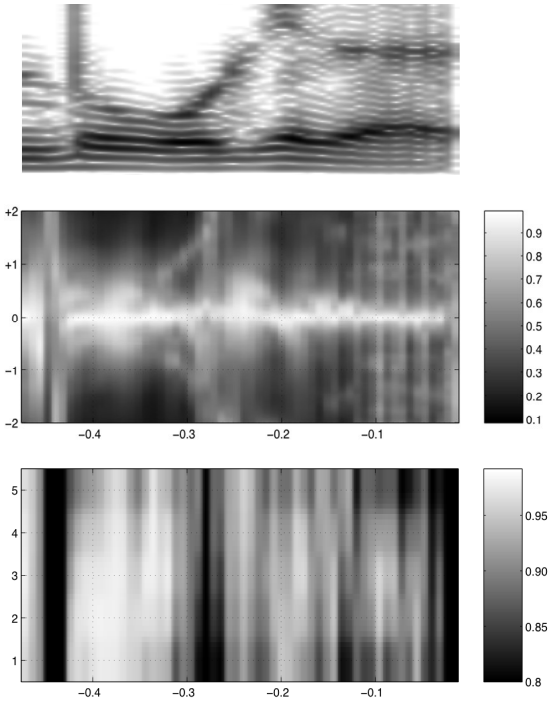

Figure 5. Spectrogram for a 500ms fragment of audio (top panel, upper frequency of 2kHz); the FFV spectrogram for the same fragment (middle panel); and the same FFV spectrum (bottom panel) after being passed through the FBNEW filterbank as shown in Figure 4.

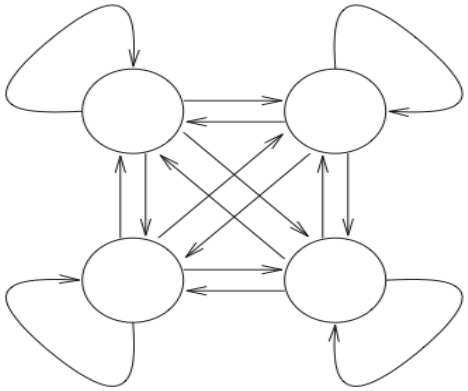

Figure 6. A fully-connected hidden Markov model (HMM) consisting of four states with one Gaussian per state.

Discussion

We have derived a continuous and instantaneous vector representation of variation in fundamental frequency and given a detailed description of the steps involved, including a graphical demonstration of both the form of the representation, and its evolution in time. We have also suggested a method for modeling sequences with HMMs and utilizing the representation in a classification task.

Initial experiments along these lines show that such HMMs, when trained on dialogue data, corroborate research on human turn-taking behavior in conversations. These experiments also suggest that the representation is suitable for direct, principled, continuous modeling (as in automatic speech recognition) of prosodic sequences, which does not require peak-identification, dynamic time warping, median filtering, landmark detection, linearization, or mean pitch estimation and subtraction (Laskowski et al., 2008a, 2008b).

We expect the method to be especially useful in situations where online processing is required, such as in conversational spoken dialogue systems. Further experiments will test the method in real systems, for example to support turn-taking decisions. We will also explore the use of the FFV spectrum in combination with other sources of information, such as durational patterns in interaction control.

Immediate next steps include fine-tuning the filter banks and the HMM topologies, and testing the results on other tasks where pitch movements are expected to play a role, such as the attitudinal coloring of short feedback utterances (e.g. Edlund, House, & Skantze, 2005; Wallers, Edlund, & Skantze, 2006), speaker verification, and automatic speech recognition for tonal languages.

Acknowledgements

We would like to thank Tanja Schultz and Rolf Carlson for encouragement of this collaboration and Anton Batliner, Rob Malkin, Rich Stern, Ashish Venugopal, and Matthias Wölfel for several occasions of discussion. The work presented here was funded in part by the Swedish Research Council (VR) project 2006-2172.

References

't Hart, J., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America, 111*(4), 1917-1930.

Edlund, J., & Heldner, M. (2005). Exploring prosody in interaction control. *Phonetica, 62*(2-4), 215-226.

Edlund, J., House, D., & Skantze, G. (2005). The effects of prosodic features on the interpretation of clarification ellipses. In *Proceedings of Interspeech 2005* (pp. 2389-2392). Lisbon, Portugal.

Laskowski, K., Edlund, J., & Heldner, M. (2008a). An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems. In *Proceedings ICASSP 2008*. Las Vegas, NV, USA.

Laskowski, K., Edlund, J., & Heldner, M. (2008b). Machine learning of prosodic sequences using the fundamental frequency variation spectrum. In *Proceedings Speech Prosody 2008*. Campinas, Brazil.

Laskowski, K., Wölfel, M., Heldner, M., & Edlund, J. (in press). Computing the fundamental frequency variation spectrum in conversational spoken dialogue systems. In *Acoustics'08 Paris*. Paris, France.

Shriberg, E., & Stolcke, A. (2004). Direct modeling of prosody: An overview of applications in automatic speech processing. In *Proceedings of Speech Prosody 2004* (pp. 575-582). Nara, Japan.

Wallers, Å., Edlund, J., & Skantze, G. (2006). The effects of prosodic features on the interpretation of synthesised backchannels. In E. André, L. Dybkjaer, W. Minker, H. Neumann & M. Weber (Eds.), *Proceedings of Perception and Interactive Technologies (PIT'06)* (pp. 183-187): Springer.