

# MODELING DYNAMIC PROSODIC VARIATION FOR SPEAKER VERIFICATION

*Kemal Sönmez*<sup>1</sup>

*Elizabeth Shriberg*<sup>1</sup>

*Larry Heck*<sup>2</sup>

*Mitchel Weintraub*<sup>1</sup>

<sup>1</sup>SRI International, Menlo Park, CA 94025

<sup>2</sup>Nuance Communications Menlo Park, CA 94025

{kemal,ees,mw}@speech.sri.com heck@nuance.com

## ABSTRACT

Statistics of frame-level pitch have recently been used in speaker recognition systems with good results [1, 2, 3]. Although they convey useful long-term information about a speaker’s distribution of  $f_0$  values, such statistics fail to capture information about local dynamics in intonation that characterize an individual’s speaking style. In this work, we take a first step toward capturing such suprasegmental patterns for automatic speaker verification. Specifically, we model the speaker’s  $f_0$  movements by fitting a piecewise linear model to the  $f_0$  track to obtain a stylized  $f_0$  contour. Parameters of the model are then used as statistical features for speaker verification. We report results on 1998 NIST speaker verification evaluation. Prosody modeling improves the verification performance of a cepstrum-based Gaussian mixture model system (as measured by a task-specific Bayes risk) by 10%.

## 1. INTRODUCTION

Statistics of frame-level pitch have recently been shown to improve the performance of state-of-the-art speaker recognition systems [1, 2, 3]. While conveying useful long-term information about a speaker’s  $f_0$  distribution, such statistics fail to capture information about local variations in intonation that constitute an individual’s speaking style (see Figure 1). Prosody is inherently suprasegmental and to make full use of intonation and/or duration information, temporal dependencies must be modeled. In this work, we take a first step toward capturing such suprasegmental patterns for automatic speaker verification. A key feature of our approach is that we use a piecewise linear stylization of the phonetic  $f_0$  contour, obtained via an automatic algorithm based on the “close-copy” stylizations detailed in [4]. Such models have been shown to generate speech that is perceptually equivalent to that produced in actual  $f_0$  contours, as verified by listening experiments. The stylization allows us to remove involuntary segmental or microintonation effects from the modeling, while retaining the fundamental patterns associated with a speaker’s intended pitch contours.

The system diagram is shown in Figure 2, dividing the task into filtering, regularization, statistical modeling and verification steps. In

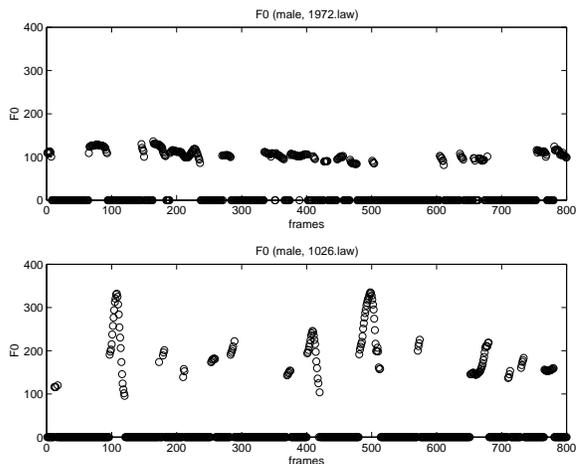


Figure 1: Difference in Intonation Style for Two Male Speakers

the filtering stage, raw pitch contours are treated for irregularities of the pitch tracker such as offshoots in the onset and end of voiced segments (note how the offshoot in the onset of the middle voiced section in Figure 3 is filtered out) and pitch halving/doubling by median filtering and robust estimation of pitch statistics by a log-normal tied mixture model [2], respectively. The regularization of the pitch contours to extract the intended intonation of a speaker is accomplished by a piecewise-linear stylization algorithm. After stylization, features that reflect the statistics of the target speaker’s habitual pitch movements are extracted from the piecewise-linear model. Parametric models are estimated for the features and used in relative entropy computation in verification.

## 2. $F_0$ CONTOUR STYLIZATION

The central idea of our approach is the regularization of  $f_0$  contours to filter out both the noise introduced by the pitch tracker and microintonation effects that hide the speaker’s intended pitch movements.

Median filtered, halving/doubling removed  $f_0$  contours still contain a lot of local variation. The so-called microintonation phe-

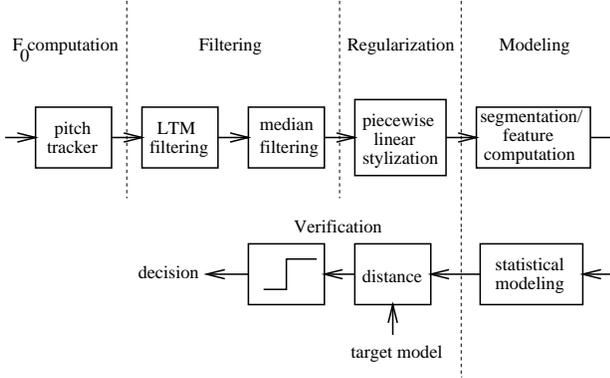


Figure 2: System Diagram

nomena are mostly involuntary on the speaker’s part and related to the incidental segmental context (i.e., phones present). We are interested in the intended intonation of the speaker (i.e., speaking style). We aim to capture this via stylization of  $f_0$  contours by a piecewise linear model. Our automatic approach to stylization of  $f_0$  contours by piecewise linear models is motivated by IPO’s close-copy stylization [4] in which the pitch tracks are modeled by linear pitch movements and the criterion for parameter estimation is perceptual indistinguishability upon reconstruction from stylized pitch tracks. Piecewise linear models, in this context, have been shown to be perceptually equivalent to real  $f_0$  contours that contain micro-intonation, as verified by listening tests.

To obtain the stylized contour, we first pre-process our  $f_0$  values to eliminate regions of pitch doubling and halving. This is achieved by applying a lognormal tied-mixture fit to the frame-level pitch values [2, 3], and filtering out pitch tracker irregularities via median filtering with a neighborhood size of 5. A piecewise linear model is then fit to the estimated (log) pitch values in voiced regions (see Figure 4). For a voiced section to be modeled by  $K$  segments, the free parameters,  $(x_k, y_k)_{k=0}^K$ , are the coordinates of the nodes that connect linear spline segments. The fitting function is given by

$$g(x) = \sum_{k=1}^K (a_k x + b_k) I_{[x_{k-1} < x \leq x_k]} \quad (1)$$

where  $(a_k, b_k)$  are the slope and the intercept of the line defined by  $(x_k, y_k)$ . The node parameters are estimated by minimizing the mean square error (MSE) between the (log) pitch estimates and the stylized fit:

$$(x_k^*, y_k^*)_{k=0}^K = \arg \min_{(x_k, y_k)_{k=0}^K} \frac{1}{T} \sum_{t=1}^T (f_0(t) - g(t))^2 \quad (2)$$

We have developed two algorithms for piecewise linear model fitting, one of which is to form a parameter vector for a given voiced section by putting together the node frame locations and log-frequency values, and using a general simplex algorithm with MSE as the objective function. In this technique, the number of nodes per voiced section is chosen proportional to the duration of the voiced

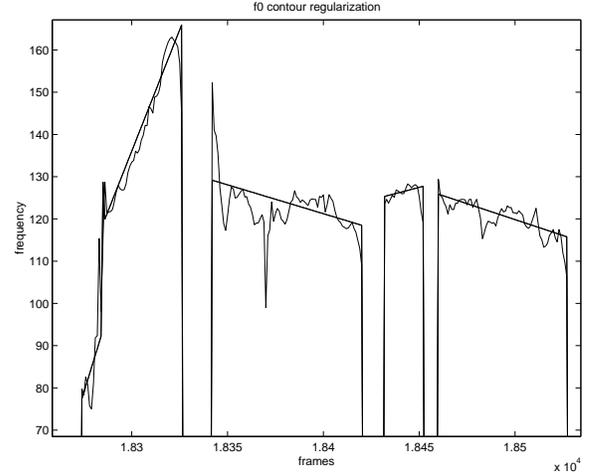


Figure 3:  $f_0$  Stylization by a Piecewise Linear Model

section. The results reported here are by using the simplex method. A newer approach we have developed estimates the extent of linear sections and introduces nodes as warranted by the local smoothness of the  $f_0$  contour.

### 3. FEATURE EXTRACTION AND STATISTICAL MODELING

The regularization of the pitch track allows us to model the statistics of the speaker’s local pitch movements. A movement is characterized by its starting and ending  $f_0$  and the rate at which the speaker realizes the movement. The stylized contours result in a significant data reduction and are used to produce the following piecewise linear model parameters as features:

- segment median:  $\log(\bar{f}_0) \sim \mathcal{N}(\mu_0, \sigma_0^2)$
- segment slope  $\bar{f}'_0 \sim \mathcal{N}(\mu_1, \sigma_1^2)$
- segment duration  $(T_s - \tau_0) \sim \mathcal{E}(\mu_2)$

which are modeled by log-normal, normal, and shifted exponential distributions, respectively. In addition to the three intonation features, we also extract the duration of continuously voiced regions, and of pauses, both modeled by shifted exponentials (see Figure 4):

- voiced segment duration  $(T_v - \tau_1) \sim \mathcal{E}(\mu_3)$
- pause duration  $(T_p - \tau_2) \sim \mathcal{E}(\mu_4)$

The scoring of the test segment against the true speaker model is accomplished in a parametric manner. Distributions are estimated for the 2-minute training segment and the 3-s, 10-s or 30-s testing segment. The negative of the relative entropy between the target and test parametric distributions is used as the score in verification.

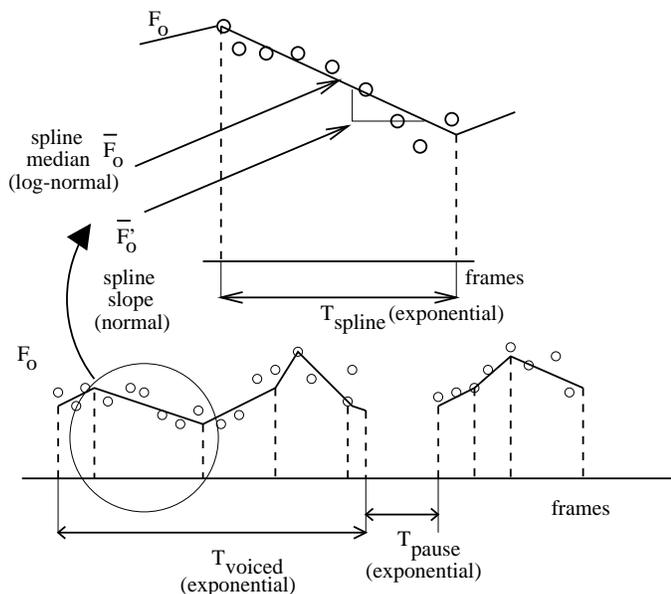


Figure 4: Stylization and Feature Extraction

#### 4. NIST SPEAKER VERIFICATION TASK

We report results on the 1998 NIST Speaker Recognition Evaluation data. The task is verification of 500 speakers (250 males, 250 females) in 5000 trials with 2 minutes of training data and 3, 10, or 30 seconds of test data over the telephone. The training condition is “1-session”, meaning the 2 minutes have come from a single telephone call.

The figure of merit for the NIST task is the detection cost function (*DCF*):

$$DCF = C_{fr}P(true)P(fr|true) + C_{fa}P(imposter)P(fa|imposter) \quad (3)$$

(i.e., Bayes Risk with priors  $P(true) = 0.01 = 1 - P(imposter)$ , and false rejection and false alarm costs  $C_{fr} = 10, C_{fa} = 1$ ). Results are intra-gender only and pooled across all target and imposter speakers.

#### 5. RESULTS

Results show that the addition of the prosodic information improved the verification performance of a state-of-the-art cepstrum-based Gaussian mixture model (GMM) system. We estimated training and testing distributions of the dynamic prosodic features, and combined the divergence between the training and testing distributions with the likelihood-ratio score of SRI’s cepstrum GMM system [3], generating a verification score that is thresholded to make the decision.

As shown in Tables 1 and 2 (and Figures 5 and 6) below, the addition of the dynamic prosodic features improves the performance

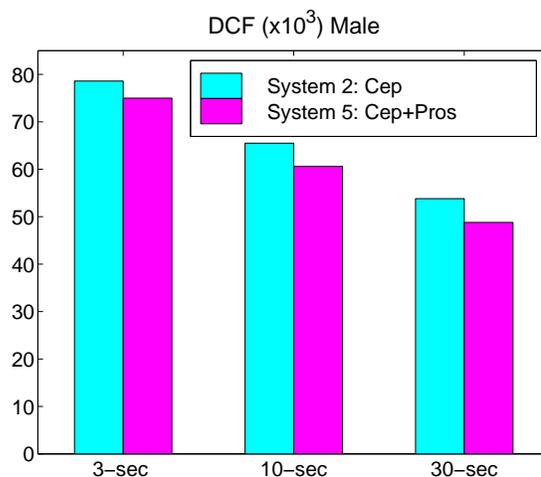


Figure 5: Detection performance for males:  $DCF(\times 10^3)$

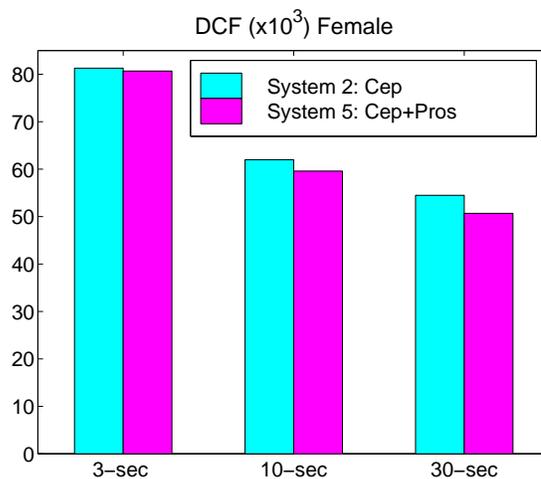


Figure 6: Detection performance for females:  $DCF(\times 10^3)$

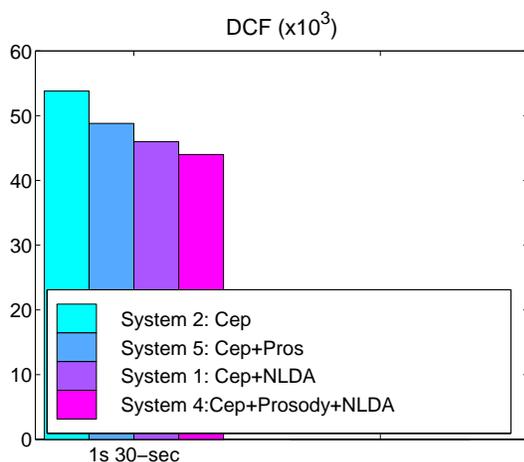
of the cepstrum GMM system significantly, by adding suprasegmental information not captured in short-term spectra. The relative improvement which, in the 3-second testing case is 4.6% and 0.7% for males and females, respectively, increases to 9.3% and 7.0% in the 30-second testing case. This is in line with the plausible expectation that for features related to dynamic variation, such as  $f_0$  contours, it takes a certain amount of data to begin to reliably distinguish individual speakers. It is possible to offer the theory that because females in general tend to have more variability in their  $f_0$  movements than males, even after normalization, we see less benefit for females than for males at the shorter test lengths. This leads to the reasonable expectation that performance for females should improve to at least the level for males if longer utterances are available.

**Table 1:** Detection performance for males:  $DCF(\times 10^3)$ 

| System             | 3-s  | 10-s | 30-s |
|--------------------|------|------|------|
| Cepstrum           | 78.6 | 65.5 | 53.8 |
| Cepstrum + prosody | 75.0 | 60.6 | 48.8 |
| Improvement        | 4.6% | 7.5% | 9.3% |

**Table 2:** Detection performance for females:  $DCF(\times 10^3)$ 

| System             | 3-s  | 10-s | 30-s |
|--------------------|------|------|------|
| Cepstrum           | 81.3 | 62.0 | 54.5 |
| Cepstrum + prosody | 80.7 | 59.6 | 50.7 |
| Improvement        | 0.7% | 3.9% | 7.0% |

**Figure 7:** Cepstrum, NLDA and prosody system combinations

The prosody scores were combined with the scores of a second system, a neural network-based system (nonlinear discriminant analysis (NLDA)) that had access to raw pitch values during training. The combined performance for males and females is shown in Figure 7. It can be observed that the prosody modeling still adds complementary information.

In summary, prosodic features have been shown to add useful complementary information to cepstral features, which results in a 7% to 10% win in DCF. Local modeling of intonation/pitch contours gives a 6.5% win. Pause and voiced segment durations give an additional 3.5% win in combination. These prosodic features are relatively inexpensive to compute and model compared with the cepstrum computation and GMM modeling.

## 6. FUTURE WORK

This work has concentrated on isolating and modeling the statistics of pitch movements. The speaking style is a function of not only the distribution of movements but also the ordering of such

movements. The introduced piecewise linear modeling regularization may provide a suitable framework for modeling more global dynamics, such as the temporal order of the splines. To this end, we have started to experiment with generating language models for pitch movements with vocabularies consisting of movement descriptions such as [baseline-to-high-with-high-slope]. N-gram models are then trained from the intonation transcriptions such a vocabulary allows. Another direction is global dynamic modeling of prosodic events by HMM modeling of  $f_0$  contours.

We have recently developed a superior piecewise linear stylization algorithm that allocates nodes according to the degree with which a given segment deviates from linearity, and is thus capable of fitting long linear segments with few nodes, and introducing more nodes in the presence of faster dynamics. This results in improved estimates for slope and segment duration and will likely improve performance overall.

## References

- [1] M.J. Carey, E.S. Parris, H. Lloyd-Thomas, and S. Bennett, "Robust Prosodic Features for Speaker Identification," in NIST Speaker Recognition Workshop, March 96.
- [2] M. K. Sönmez, L. Heck, M. Weintraub, and E. Shriberg, "A lognormal tied mixture model of pitch for prosody-based speaker recognition," in EUROSPEECH '97, vol. 3, (Rhodes, Greece), pp. 1391–1394, September 1997.
- [3] M. K. Sönmez, L. Heck, and M. Weintraub, "Prosodic Features for Speaker Recognition," in NIST Speaker Recognition Workshop, April, 1998.
- [4] J. 't Hart, R. Collier, and A. Cohen, "A Perceptual Study of Intonation," Cambridge University Press, 1990, Cambridge, UK.
- [5] Y. Konig, L. Heck, M. Weintraub, and K. Sönmez "Nonlinear Discriminant Feature Extraction for Robust Text-Independent Speaker Recognition," Proc. of RLA2C - Speaker Recognition and Its Commercial and Forensic Applications, 1998, Avignon, France.