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

Problem

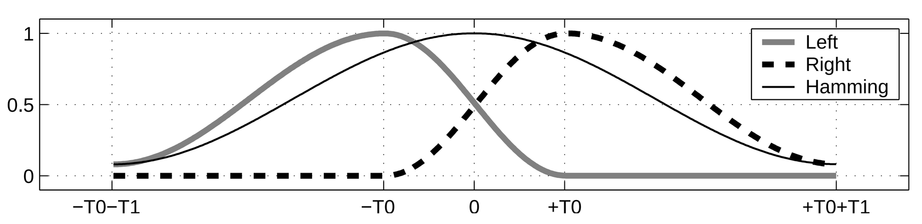
Current state-of-the-art conversational spoken dialogue systems are not sufficiently responsive. They produce speech at detected end-of-utterance (EOU) locations; EOU detection consists of **waiting for 0.5–2.0 seconds**.

Goal

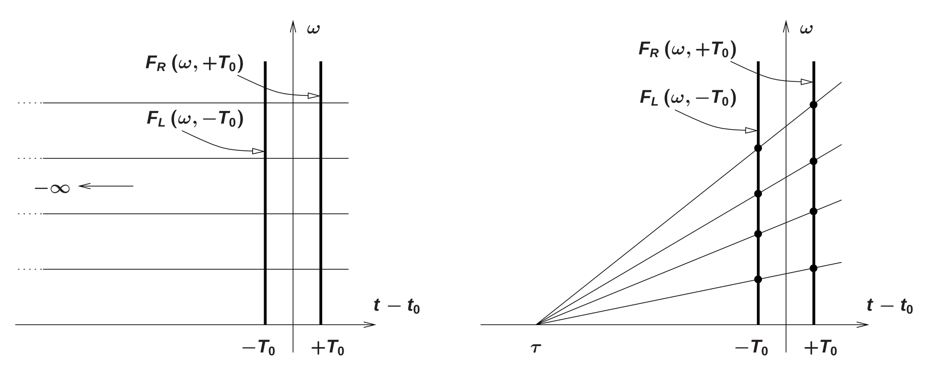
Faster online prediction ($\lesssim$0.3 s) of end-of-utterance (EOU) locations.

The Fundamental Frequency Variation Spectrum

- use **entire** spectrum to quantify variation in F0
- sample spectrum at two locations in each frame: $-T_0$ and $+T_0$ relative to midpoint of frame



- define the **vanishing-point product** at a vanishing point τ



$$g^{\tau}(\tau) = \begin{cases} \int_{-f_s/2}^{+f_s/2} F_L\left(\left(\frac{-\tau-T_0}{-\tau+T_0}\right) f\right) F_R^*(f)\, df, & \tau < -T_0 \\ \int_{-f_s/2}^{+f_s/2} F_L(f)\, F_R^*\left(\left(\frac{+\tau-T_0}{+\tau+T_0}\right) f\right) df, & \tau > +T_0 \end{cases}$$

- define the conformal mapping

$$\rho = \begin{cases} -\log_2\left(\frac{-\tau-T_0}{-\tau+T_0}\right), & \tau < -T_0 \\ +\log_2\left(\frac{+\tau-T_0}{+\tau+T_0}\right), & \tau > +T_0 \end{cases}$$

and transform $g^{\tau}(\tau)$ to yield

$$g^{\rho}(\rho) = \begin{cases} \int_{-f_s/2}^{+f_s/2} F_L(f)\, F_R^*\left(2^{+\rho} f\right) df, & \rho < 0 \\ \int_{-f_s/2}^{+f_s/2} F_L\left(2^{-\rho} f\right) F_R^*(f)\, df, & \rho \geq 0 \end{cases}$$

- define the linear interpolation

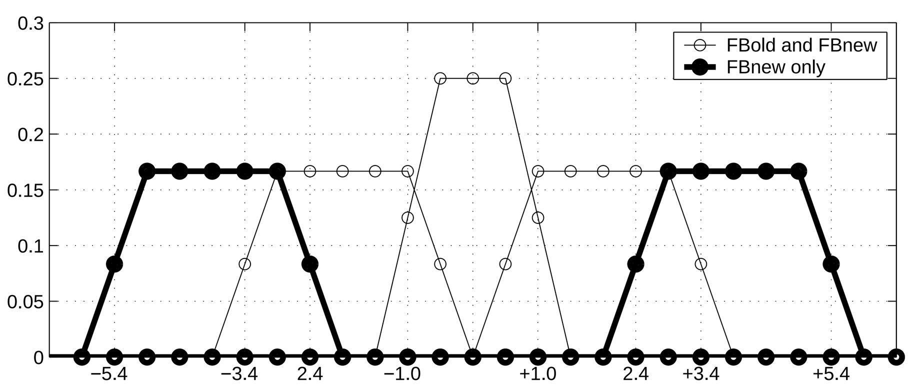
$$|\tilde{F}\left(2^{\pm\rho}k\right)| = \beta\, |F\left[\lfloor 2^{\pm\rho}k \rfloor\right]| + (1-\beta)\, |F\left[\lceil 2^{\pm\rho}k \rceil\right]|$$

where $\beta = \left|\lceil 2^{\pm\rho}k \rceil - 2^{\pm\rho}k\right|$

- sample $g^{\rho}(\rho)$ at discrete locations to yield

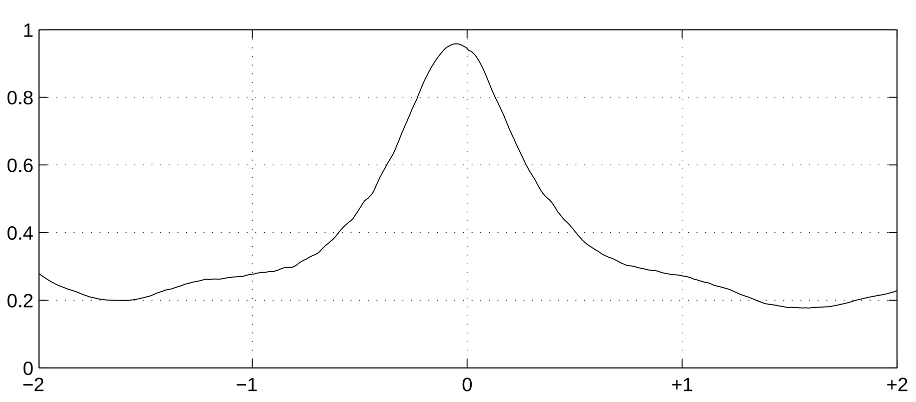
$$g^{\rho}[r] = \begin{cases} \sum_{k=-N/2+1}^{N/2} |\tilde{F}_L\left(2^{-4r/N}k\right)|\, |F_R^*[k]|, & r \geq 0 \\ \sum_{k=-N/2+1}^{N/2} |F_L[k]|\, |\tilde{F}_R^*\left(2^{+4r/N}k\right)|, & r < 0 \end{cases}$$

- normalize for energy-indepenendence, and apply filterbank

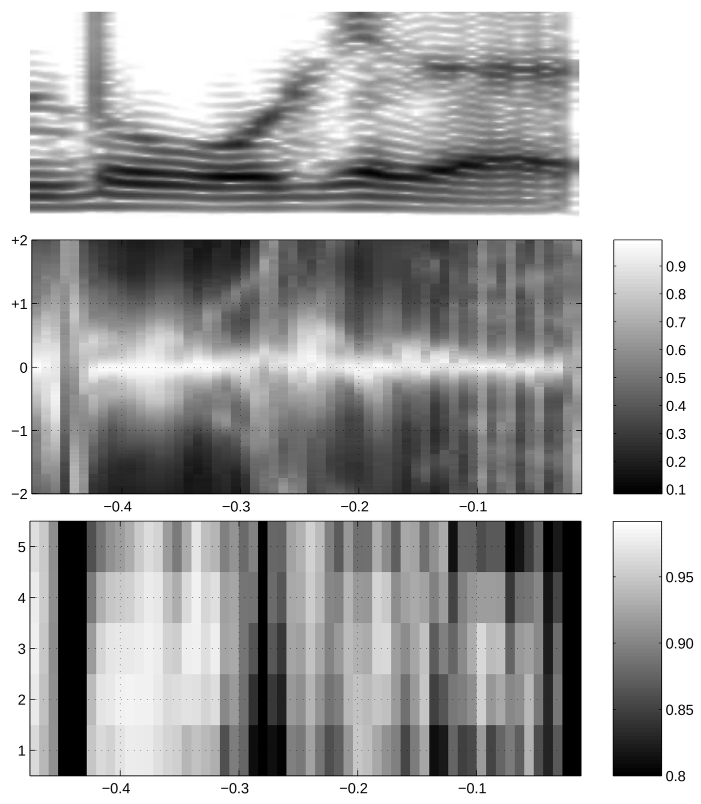


- apply Karhunen-Loéve whitening transform

Sample "Spectrum" Representation



Sample "Spectrogram" Representation



Modeling

Hidden Markov model for each of **SC** and **¬SC**:

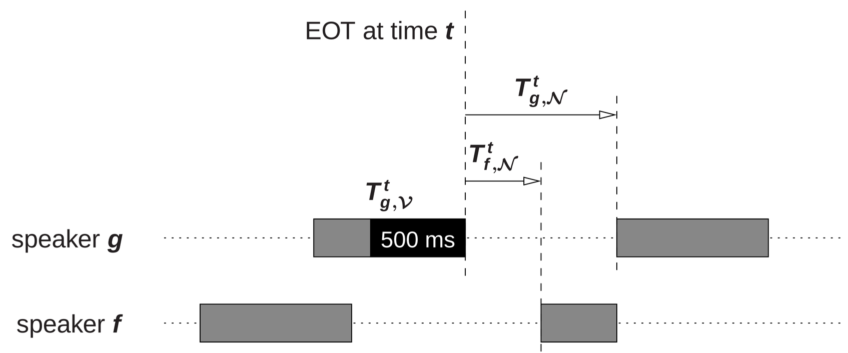
- 4 states
- 1 Gaussian per state
- trained on DEVSET using the Forward-Backward Algorithm

Online Speaker Change Prediction

Binary classification of each end-of-talkspurt (EOT) location in a human-human dialogue as either

- *a speaker change*, **SC**; or as
- *not a speaker change*, **¬SC**

using a **prosodic description of 500 ms of audio** preceeding the EOT.



Reference labels are given by the **automatic assignment**:

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

In previous work, we have demonstrated that the occurrence of *observed* speaker changes at time t is strongly correlated with human judgement that they are *appropriate*.

Data

Swedish Map Task Corpus:

- two speakers: a *giver*, g, and a *follower*, f
- task: g explains directions to f

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

- highly interactive dialogues
- DEVSET and EVALSET are disjoint in speakers

Automatic Classification

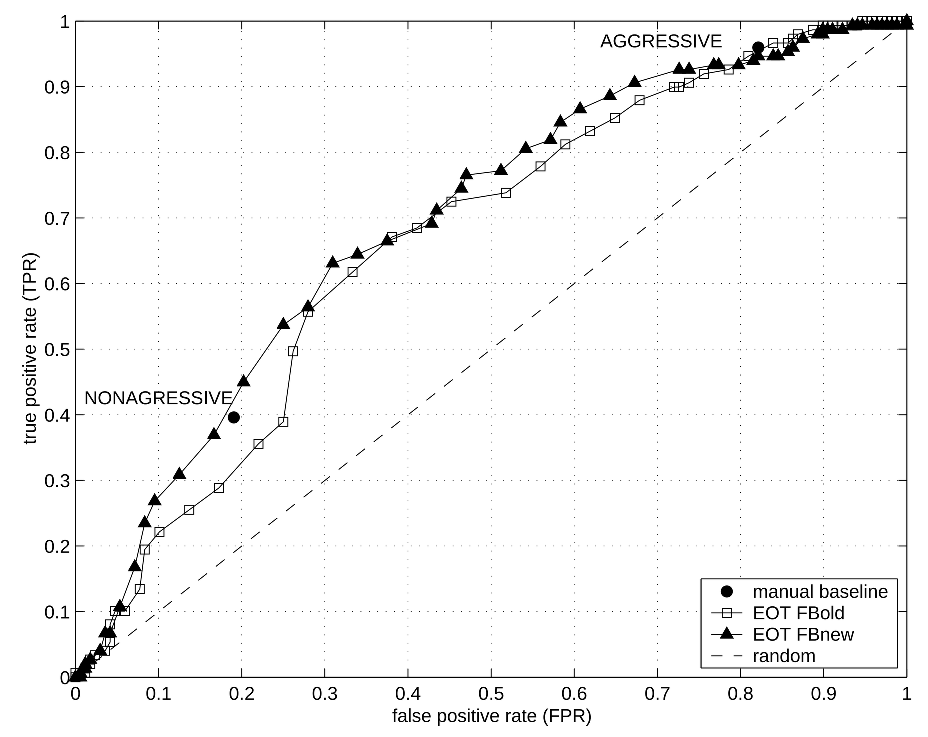
- compare two different log-likelihood-ratio classifiers
- train 10 HMMs for each of SC and ¬SC, by using different random seeds prior to Forward-Backward training
- **classifier 1**: log-likelihood-ratio over **mean** of 10 model likelihoods

$$L_t = \arg\max_k P\left(\mathbf{x} \,|\, \mathcal{M}_k\right) \quad (1)$$

- **classifier 2**: log-likelihood-ratio over **product** of 10 model likelihoods

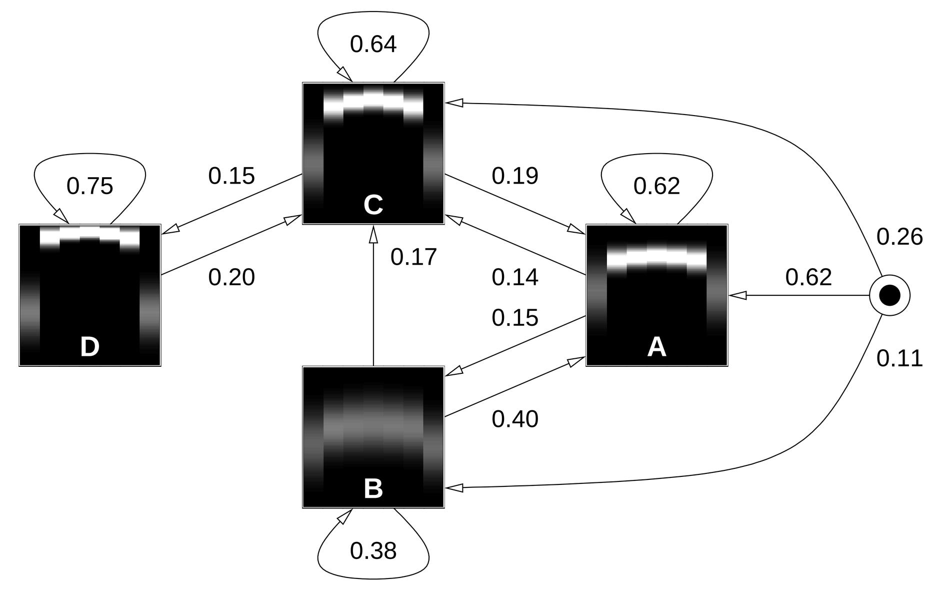
$$L_t = \arg\max_k \prod_{i=1}^{10} P\left(\mathbf{x} \,|\, \mathcal{M}_{k,i}\right) \quad (2)$$

Experimental Results

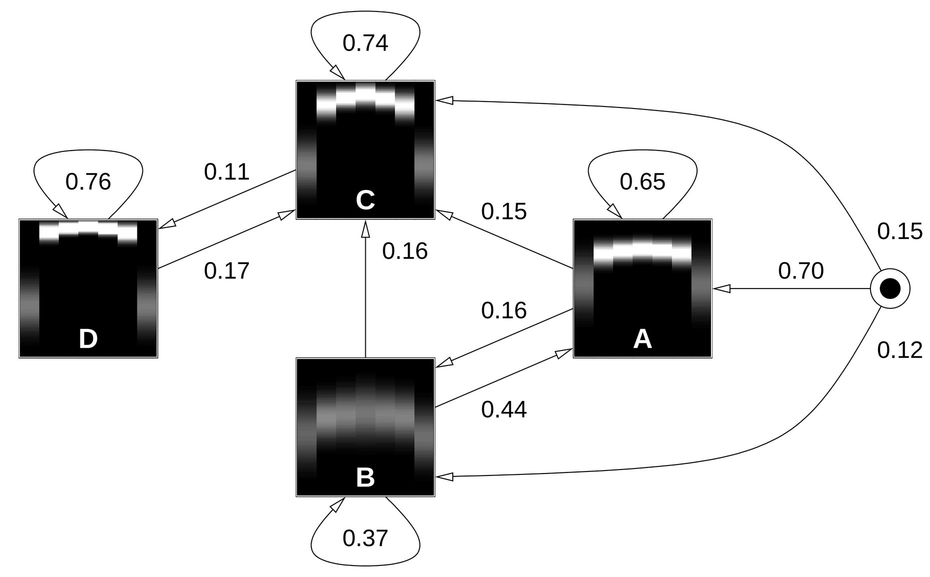


Learned Sequences

- randomly chosen HMMs, showing transition and (unnormalized) emmission probabilities:

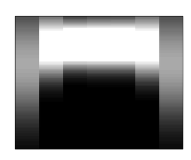


SC

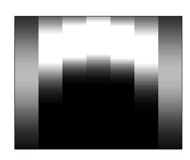


¬SC

- normalized emmission probabilities for state C in both models:



State C for SC



State C for ¬SC

Conclusions

- first exploration of what models of fundamental frequency variation sequences actually learn
- learned models corroborate existing research of human behavior
- represenation appears suitable for direct, principled, continuous sequence modeling as in SAD and ASR
- improved filterbank design yields SC/¬SC classification accuracy improvements on unseen data of 12-17% relative

kornel@cs.cmu.edu, {edlund,heldner}@speech.kth.se