The fundamental frequency variation spectrum

Kornel Laskowski
interACT, Carnegie Mellon University

Mattias Heldner & Jens Edlund
Centre for Speech Technology, KTH

1

Purpose

To describe a new vector representation of variation in fundamental frequency – the fundamental frequency variation spectrum – which has a number of desirable properties for use in online speech processing, e.g. in spoken dialogue systems

2

Background: pitch tracking

- ASR has long ago switched from formant peak localization to spectral formant representations covering the whole spectrum
- Prosodic processing continues to rely on localization of a peak corresponding to the first harmonic (F0)

Background: pitch tracking

- Peak localization in acoustic signals is error prone, and pitch trackers employ a number of techniques to improve robustness
- The improved robustness comes at a cost (dependency on right context, delayed F0 estimates, biased F0 distributions...) making such F0 estimates less suited for online processing tasks, e.g. in spoken dialogue systems

4

Background: F0 modeling

- What's more, prosodic modeling if often concerned with **F0 variation** – e.g. flat, falling and rising pitch patterns – requiring additional processing steps (e.g. stylization, curve fitting, regression, normalization, semitone transformation...) introducing additional processing delays
- See e.g. Edlund, J., & Heldner, M. (2005). Exploring prosody in interaction control. *Phonetica*, 62(2-4), 215-226 for examples of this...

5

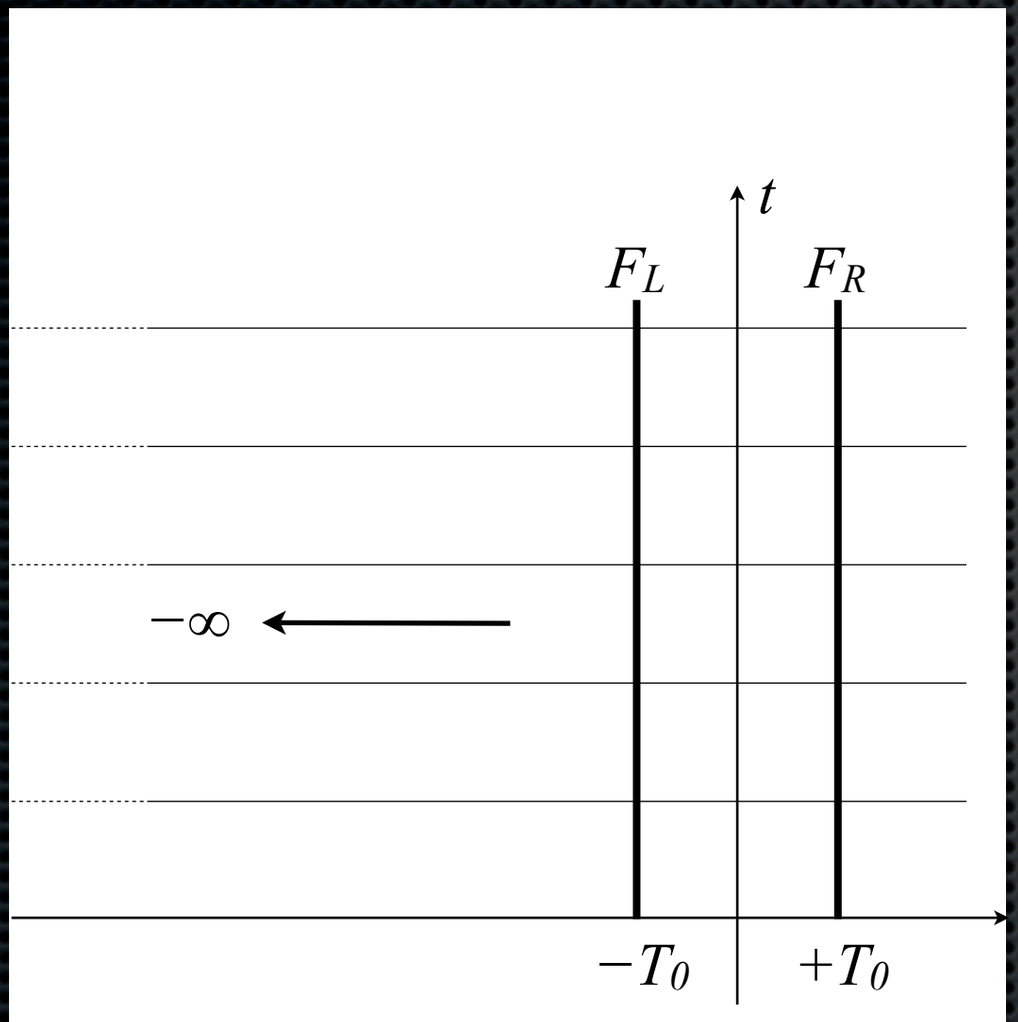
The FFV spectrum is

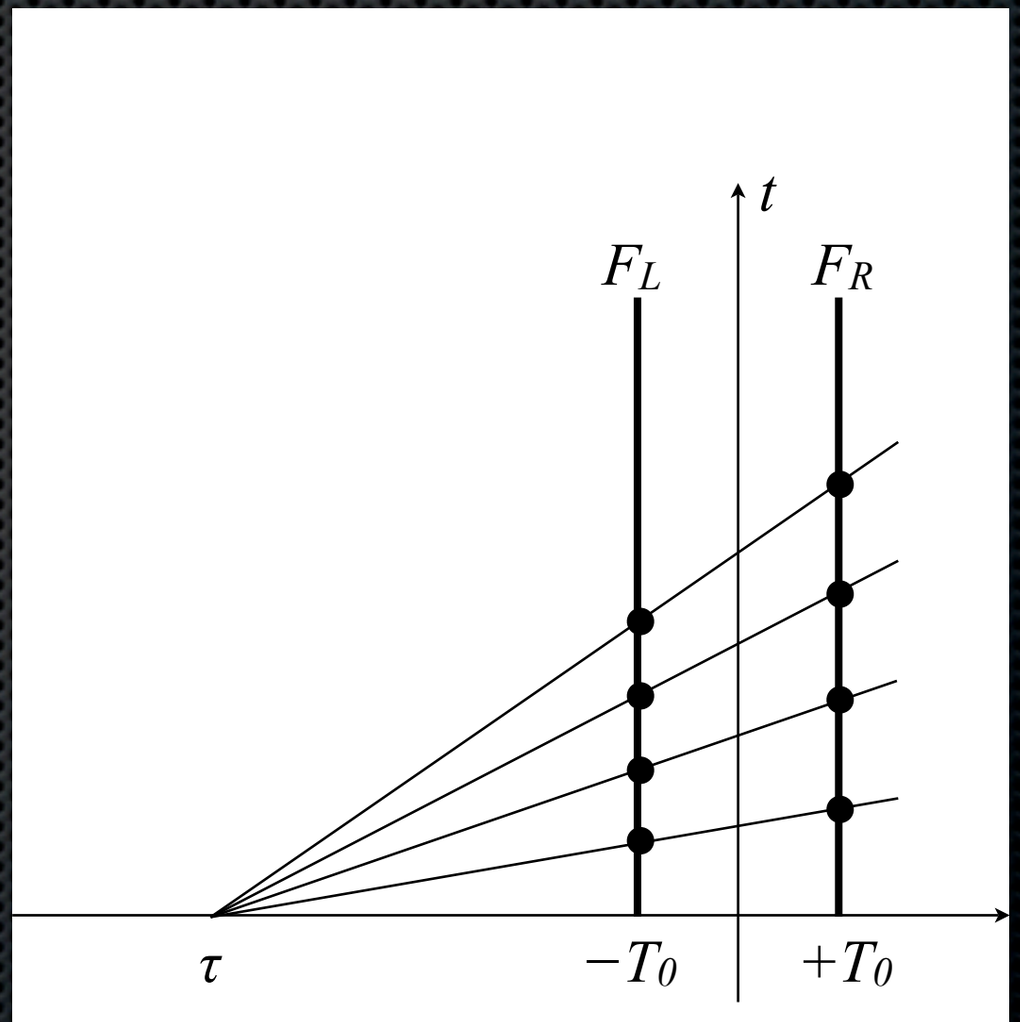
1. instantaneous (not relying on adjacent frames or right context)
2. continuous (defined for all frames, incl. voiceless ones)
3. distributed (no statistical bias introduced)
4. potentially sparse (suitable for standard acoustic modeling techniques)

useful for modeling prosodic sequences in online processing

6

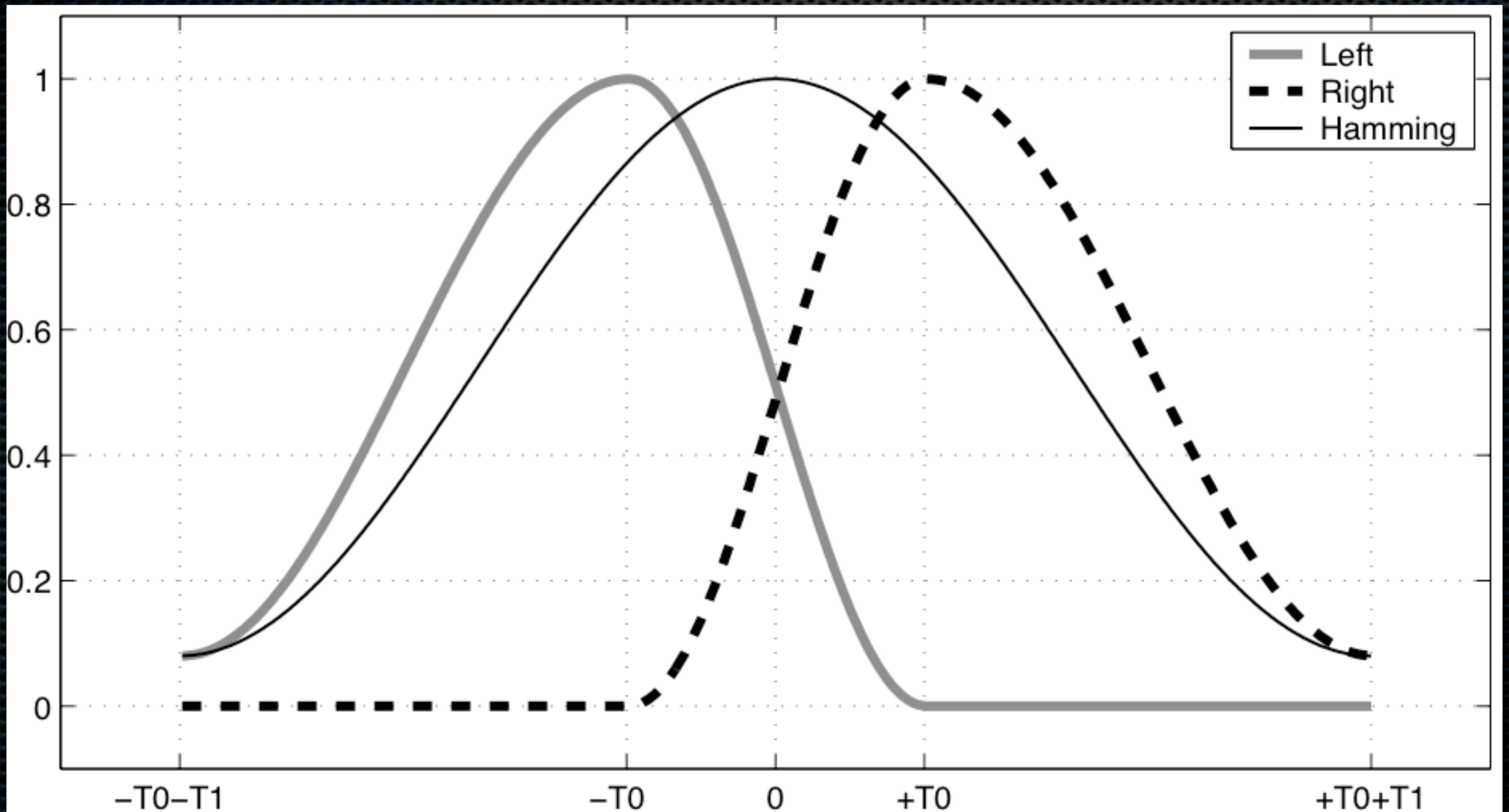
Vanishing-point perspective dot product





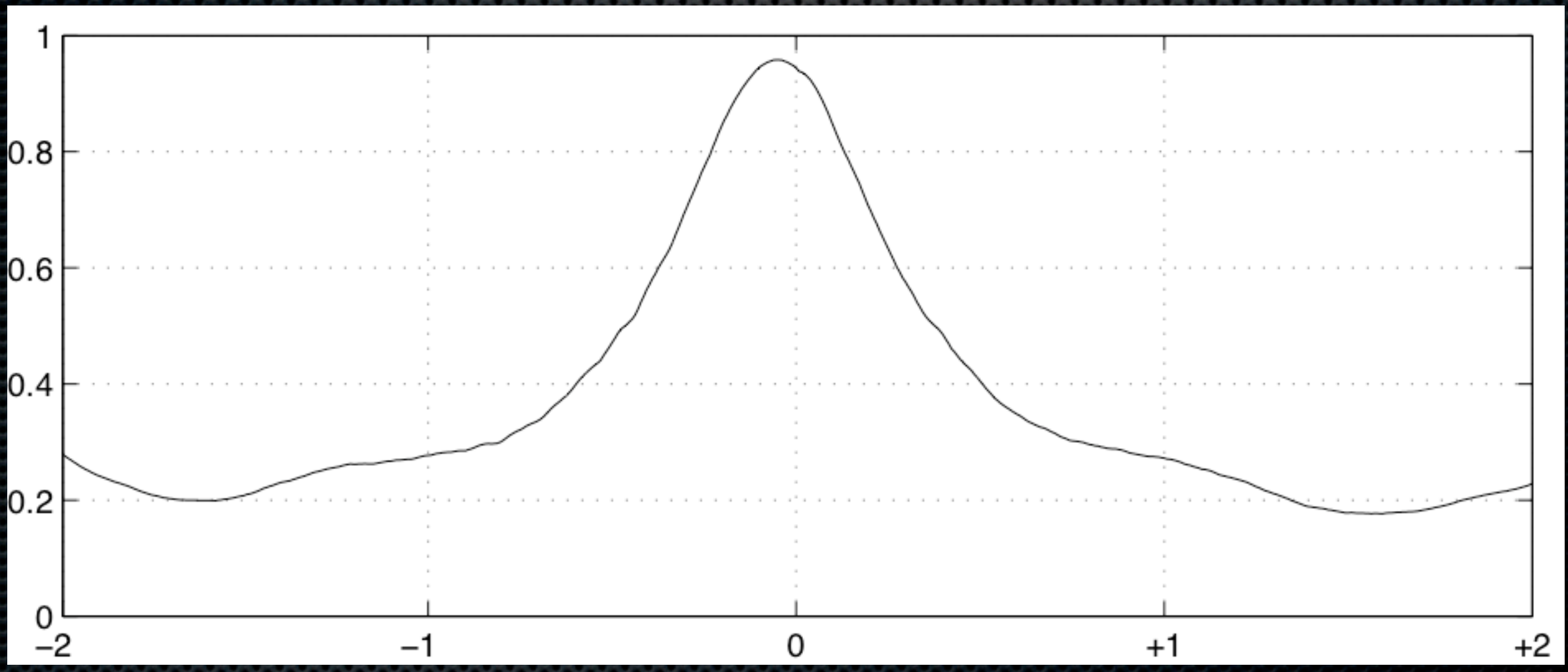
7

Windows used for computation of F_L and F_R



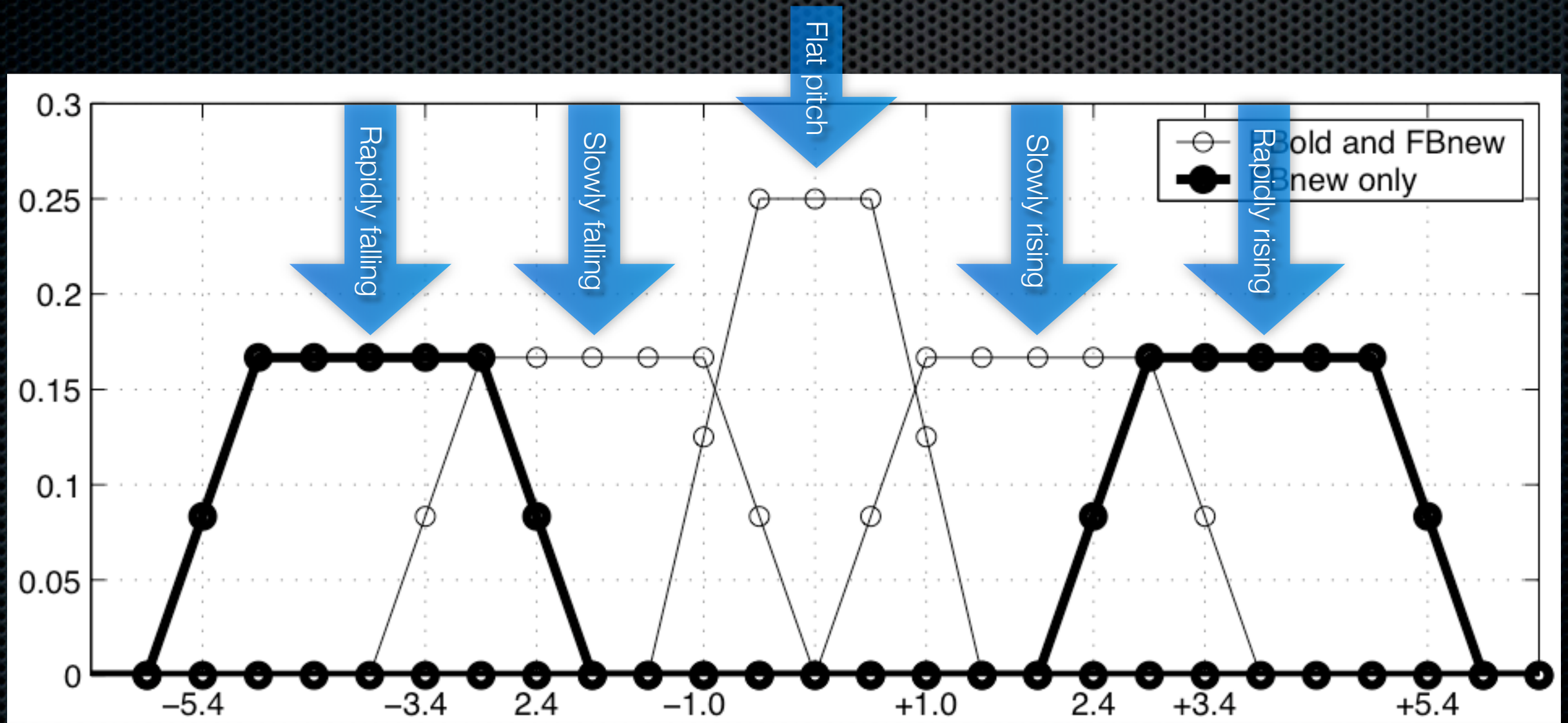
8

Sample FFV spectrum



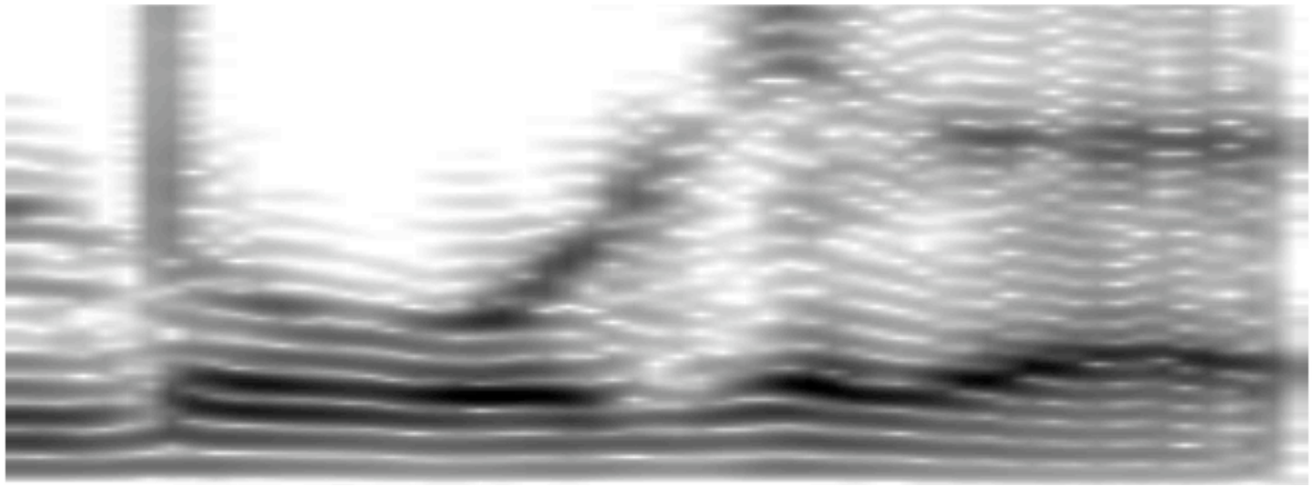
9

Filterbank

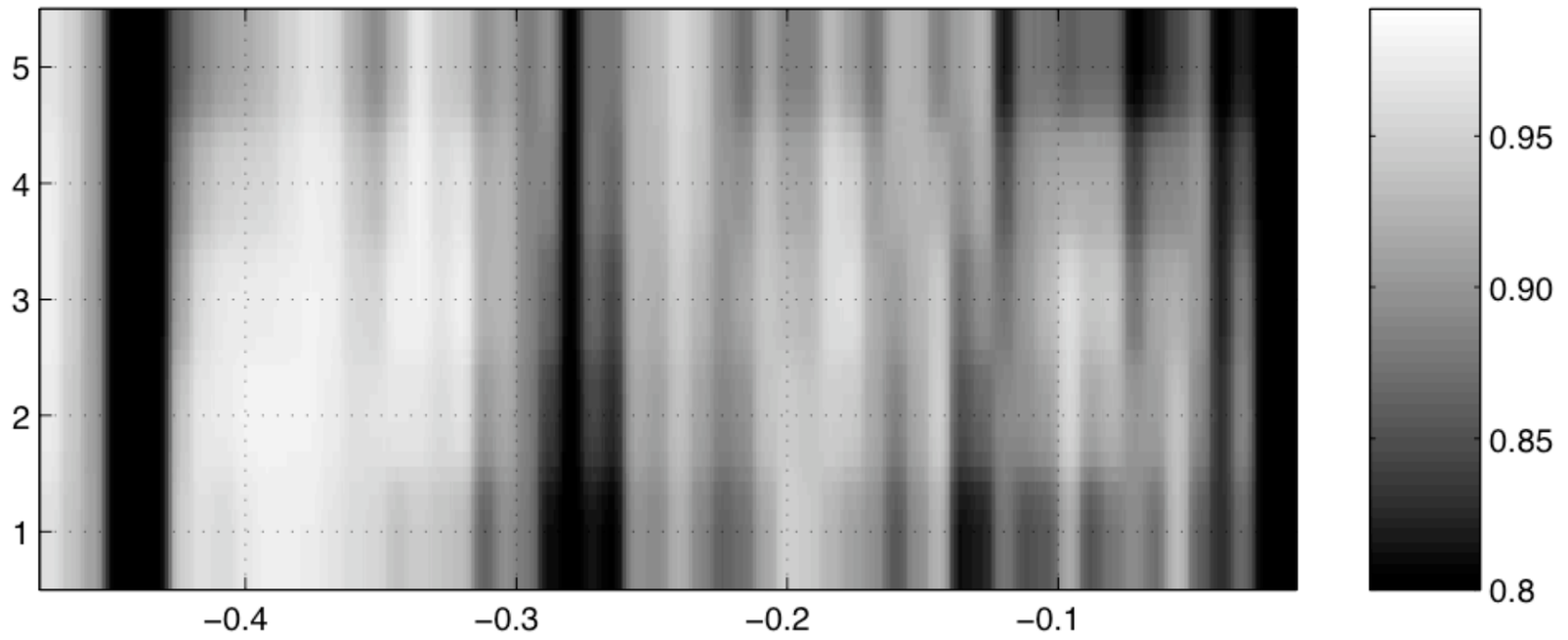


10

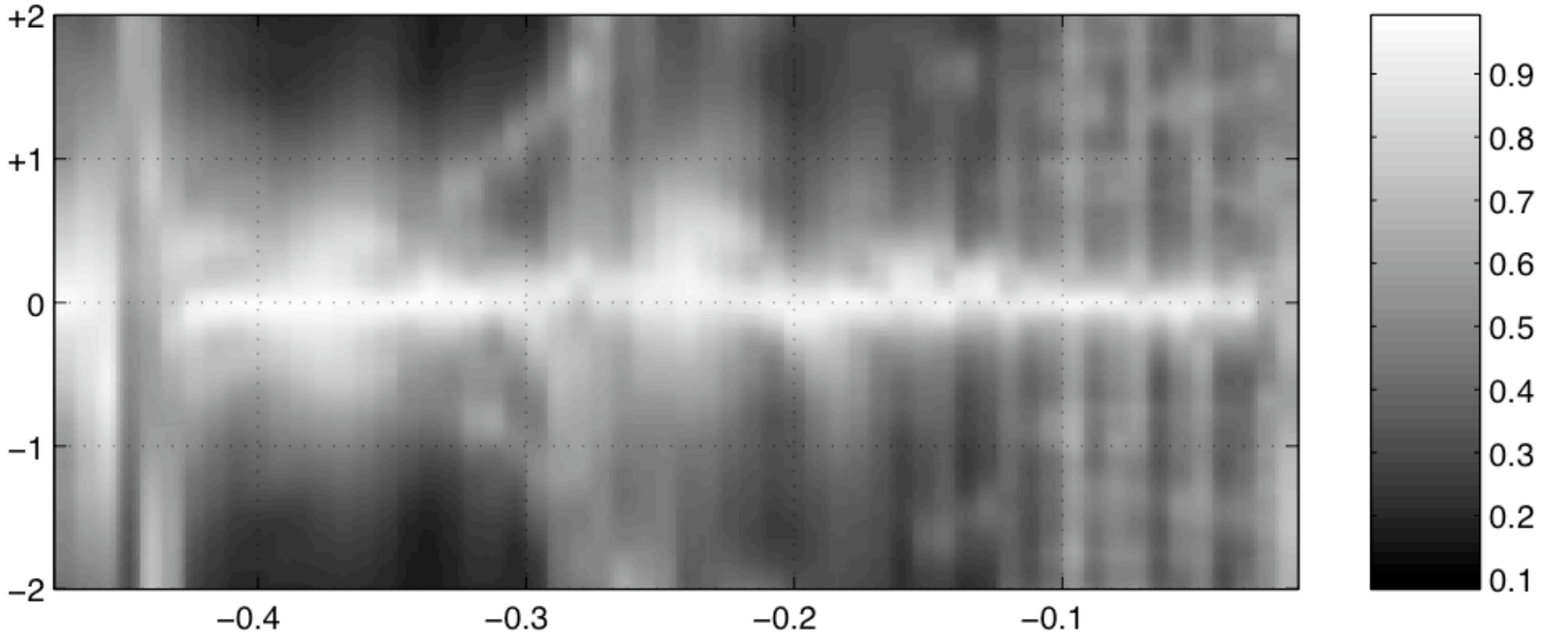
Spectrogram



FFV spectrum



FFV spectrum passed through filterbank



Modeling sequences of FFV spectra

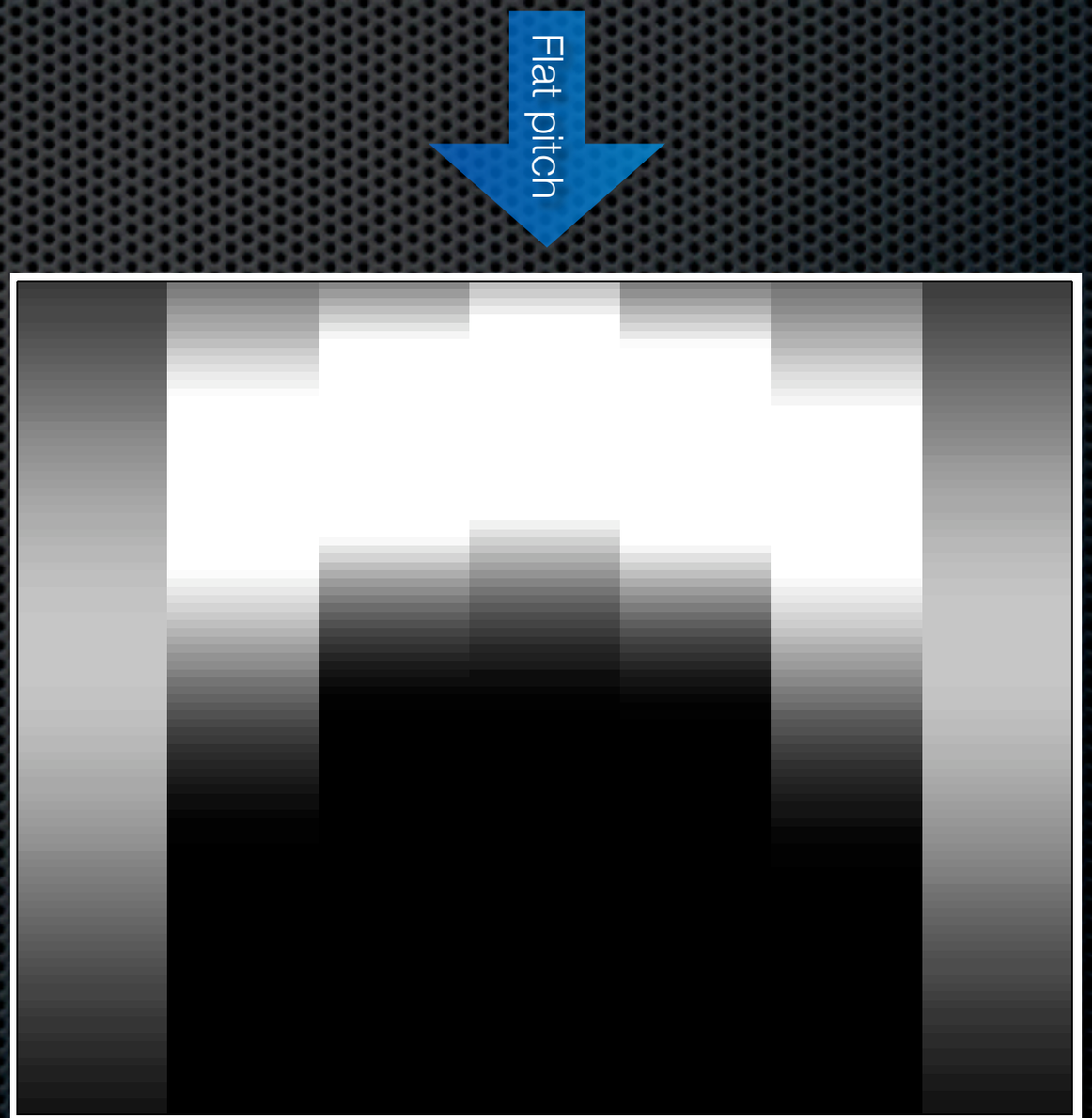
- How to arrive at something more like what we normally associate with prosody?
  - From FFV spectra passed through a filterbank to e.g. flat, falling and rising pitch movements?
- Standard option for modeling sequences: training HMMs
- We have used fully-connected HMMs with four states and one gaussian per state

12

Modeling example

- **Task**: Classification of speaker changes vs. speaker holds
- **Prediction**: Flat pitch before silence is often found in speaker holds

Speaker hold model dominated by activity in the perceptually flat filter



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

- Laskowski, K., Edlund, J., & Heldner, M. (2008). Machine learning of prosodic sequences using the fundamental frequency variation spectrum. In ***Proceedings of the Speech Prosody 2008 Conference*** (pp. 151-154). Campinas, Brazil: Editora RG/CNPq.

- Laskowski, K., Edlund, J., & Heldner, M. (2008). An instantaneous vector representation of delta pitch for speaker-change prediction in conversational dialogue systems. In ***Proceedings ICASSP 2008*** (pp. 5041-5044). Las Vegas, Nevada, USA.

14

Thank you for your attention

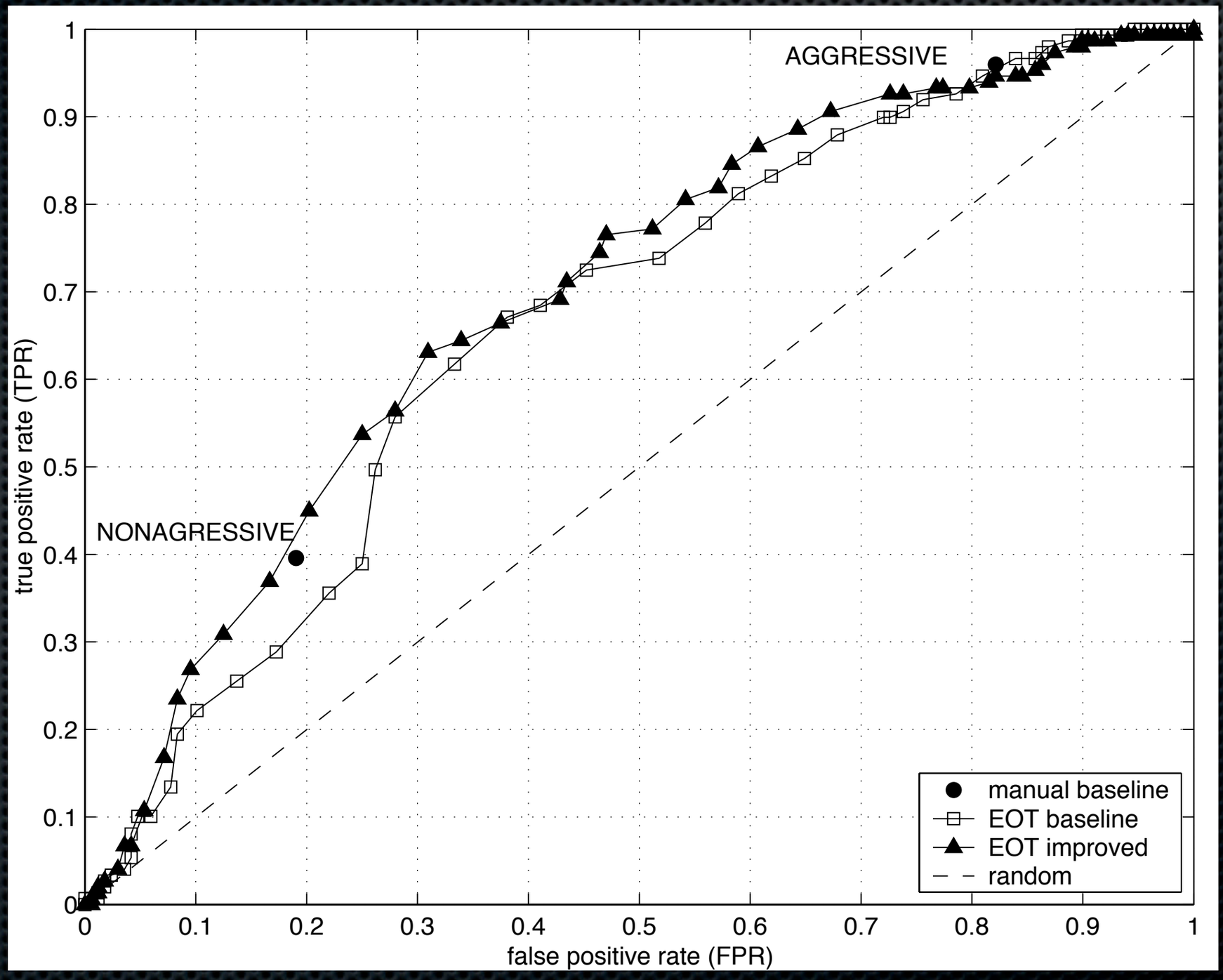
Work funded in part by the Swedish Research Council (VR) project 2006-2172

Vad gör tal till samtal?

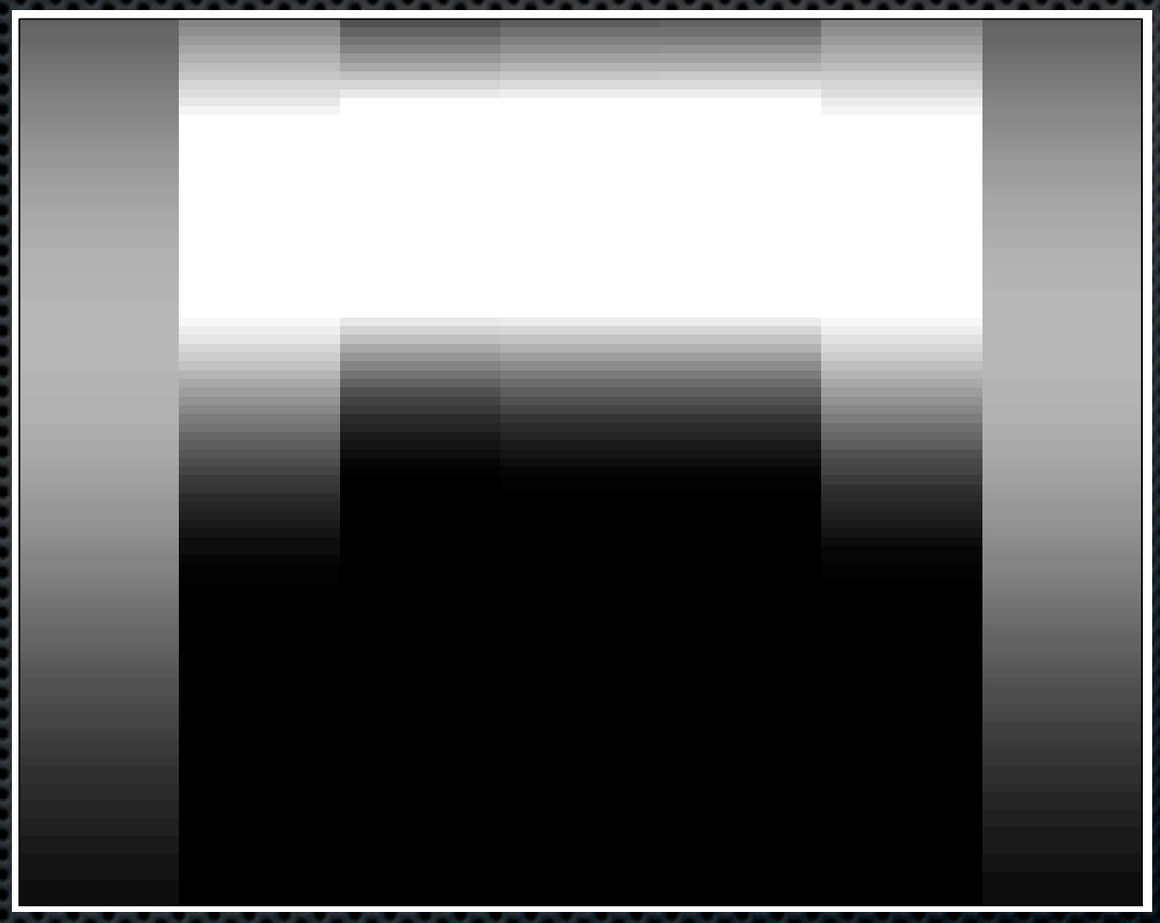
www.speech.kth.se/vats

15

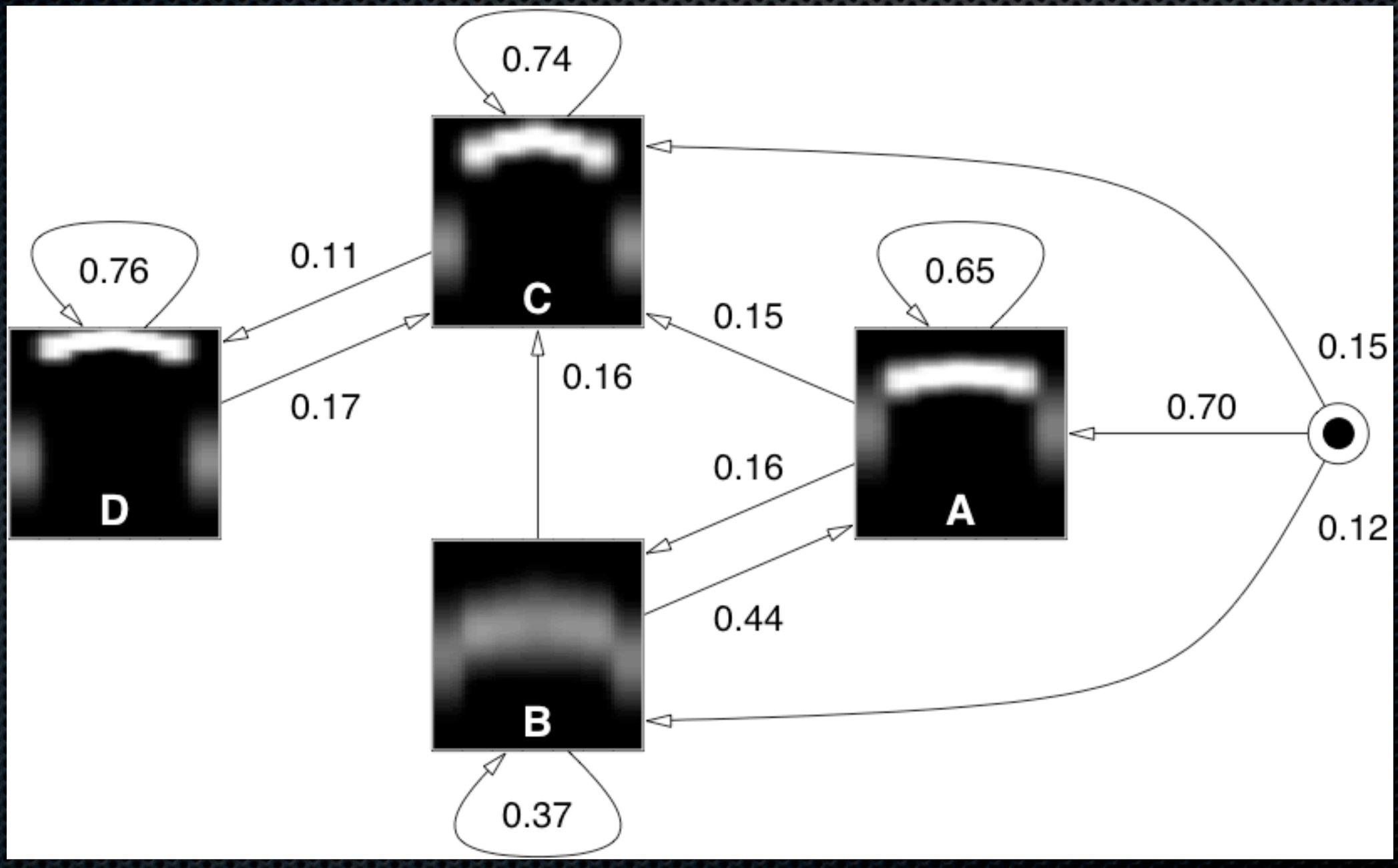
ROC for speaker change vs. speaker hold classification



Speaker change model



Speaker hold model



Speaker change model