An Instantaneous Vector Representation of Delta Pitch for Speaker-Change Prediction in Conversational Dialogue Systems

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 1–2 seconds**.

Goal

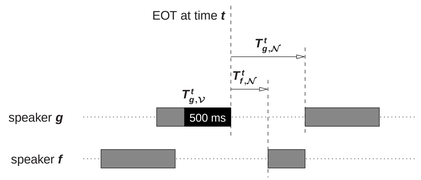
Faster prediction (0.3 s) of end-of-utterance (EOU) locations.

Approach

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

- *a speaker change*, $\mathbf{SC}$; or as
- *not a speaker change*, $\neg\mathbf{SC}$

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



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

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

Data

Swedish Map Task Corpus:

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

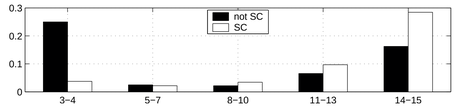
Data Set	Duration	Dialogue role ***g***		
	(mn:ss)	speakers	# EOTs	# SCs
DevSet	77:40	F4,F5,M2,M3	480	222
EvalSet	60:39	F1,F2,F3,M1	317	149

- highly interactive dialogues
- DevSet and EvalSet are disjoint in speakers

Validation of Approach

Question: Are the **observed**, automatically labeled $\mathbf{SC}$ events actually deemed **appropriate** by human judges?

Answer: Our analysis shows that, for the most part, **yes**:



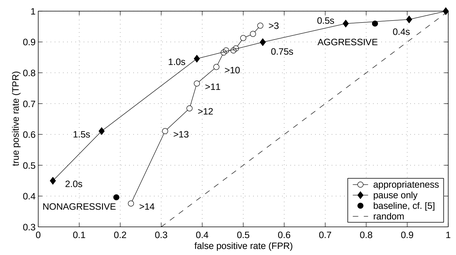
Caveats:

- Many more opportunities for ***f*** to begin speaking are deemed appropriate than are actually made use of by ***f***: $\mathbf{SC}$ prediction inherently leads to **many false alarms.**
- Some opportunities taken by ***f*** to begin speaking are deemed inappropriate: $\mathbf{SC}$ prediction inherently leads to **some misses**.
- Both aspects are due to **uncertainty** in highly interactive human-human dialogue.

Baseline Performance

Three contrastive systems:

- **human appropriateness system**: declare each EOT as an SC if summed Likert scores from 3 human judges exceed a fixed threshold
- **pause-only system**: declare each EOT as an SC if post-EOT pause duration exceeds a fixed threshold
- **rule-based F0 system**: declare each EOT as an SC if F0 slope and range features meat hand-crafted criteria



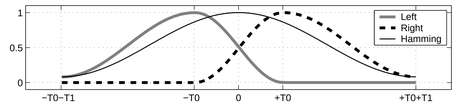
Some Desirable Properties of a Prosodic Reprentation

Would like a frame-level representation of F0 variation which is:

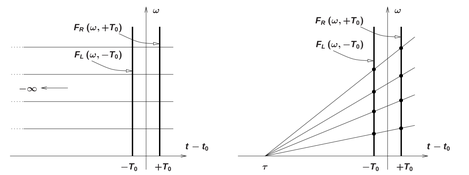
- **instantaneous**: not relying on adjacent frames
- **continuous**: defined for all time
- **distributed**: not relying on single harmonic peak location
- **sparse**: suitable for application of standard acoustic modeling techniques

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

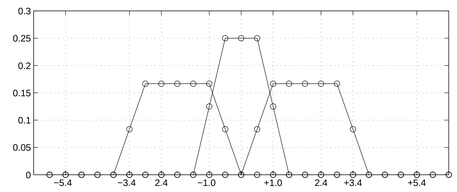
$$|\tilde{F}\left(2^{\pm\rho}k\right)| = \beta\,|F\left[\lfloor 2^{\pm\rho}k \rfloor\right]| + (1-\beta)\,|F\left[\lceil 2^{\pm\rho}k \rceil\right]|$$

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



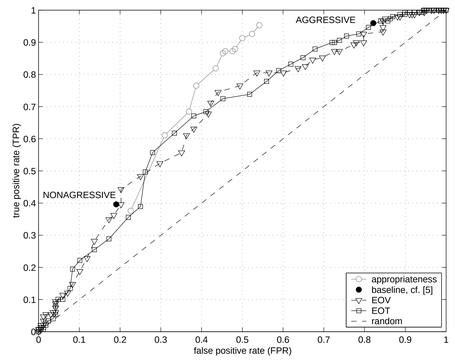
- apply Karhunen-Loéve whitening transform

Acoustic Model

Hidden Markov model for each of $\mathbf{SC}$ and $\neg\mathbf{SC}$:

- 4 states
- 1 Gaussian per state
- trained on DevSet using the Forward-Backward Algorithm
- classify EOTs using log-likelihood comparison

Experimental Results



Conclusions

- derivation of a continuous vector-valued representation of instantaneous F0 variation
- representation is compatible with standard acoustic modeling techniques
- fully automatic labeling framework allows for inexpensive training with large amounts of data
- models successfully discard inappropriate locations to speak, in spite of label mismatch
- new system outperforms pause-only systems
- new system reproduces baseline hand-crafted pitch-based system performance to within 1%, without relying on pitch range

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