Learning Prosodic Sequences Using the Fundamental Frequency Variation Spectrum

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 (<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: \(-\tau_0\) and \(+\tau_0\) relative to midpoint of frame

\[ g^T(\tau) = \begin{cases} \frac{1}{\tau_0^2} \int_{-\tau_0}^{\tau} F_k(\tau) F_T(f) \ df, & \tau < -\tau_0 \\ \frac{1}{\tau_0^2} \int_{-\tau_0}^{\tau} F_k(\tau) F_T(f) \ df, & \tau > +\tau_0 \\ \end{cases} \]

- define the vanishing-point product at a vanishing point \(\tau\)

\[ \rho = \begin{cases} -\log_2 \left( \frac{\tau}{\tau_0^2} \right), & \tau < -\tau_0 \\ +\log_2 \left( \frac{\tau}{\tau_0^2} \right), & \tau > +\tau_0 \\ \end{cases} \]

and transform \(g^T(\tau)\) to yield

\[ g^\rho(\rho) = \begin{cases} \frac{1}{\tau_0^2} \int_{-\tau_0}^{\tau} F_k(\tau) F_T(f) \ df, & \rho < 0 \\ \frac{1}{\tau_0^2} \int_{-\tau_0}^{\tau} F_k(\tau) F_T(f) \ df, & \rho \geq 0 \\ \end{cases} \]

- define the linear interpolation

\[ | F \left( 2^\rho / k \right) | = \beta | F \left( 2^{\rho-1} / k \right) | + (1 - \beta) | F \left( 2^{\rho} / k \right) | \]

where \(\beta = \frac{2^{\rho-1} / k - 2^\rho / k}{2^\rho / k - 2^{\rho-1} / k}\)

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

\[ g^\rho(\rho) = \begin{cases} \frac{1}{\tau_0^2} \int_{-\tau_0}^{\tau} F_k(\tau) F_T(f) \ df, & \rho \geq 0 \\ \frac{1}{\tau_0^2} \int_{-\tau_0}^{\tau} F_k(\tau) F_T(f) \ df, & \rho < 0 \\ \end{cases} \]

- normalize for energy-independence, and apply filterbank

- apply Karhunen-Loève whitening transform

Sample “Spectrum” 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 preceding the EOT.

Reference labels are given by the automatic assignment:

\[ L_T = \begin{cases} SC \quad & \text{if } T_{F_T}^{\text{SC}} - T_{F_T}^{\text{¬SC}} < 0 \\ ¬SC, \ otherwise \end{cases} \]

In previous work, we have demonstrated that the occurrence of observed speaker changes at time \(T\) is strongly correlated with human judgment that they are appropriate.

Data

Swedish Map Task Corpus:

- two speakers: a giver, \(g\), and a follower, \(f\)
- task: \(g\) explains directions to \(f\)

- manually baseline

- 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

\[ L_1 = \arg \max_k P(x | X_k) \quad (1) \]

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

\[ L_2 = \arg \max_k \prod_{j=1}^{10} P(x | X_k) \quad (2) \]

Conclusions

- first exploration of what models of fundamental frequency variation sequences actually learn
- learned models corroborate existing research of human behavior
- representation 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