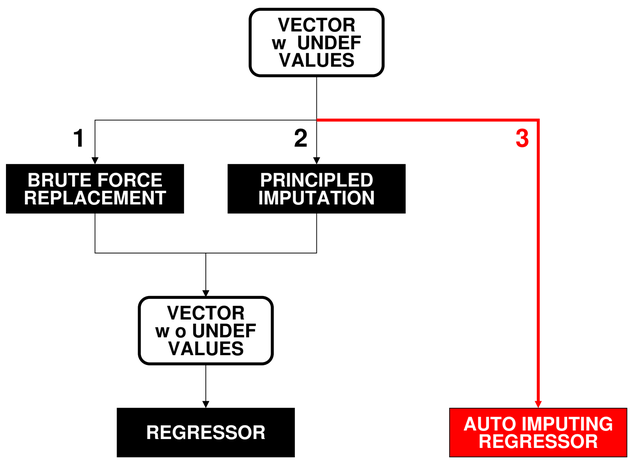
AUTO-IMPUTING RADIAL BASIS FUNCTIONS FOR NEURAL-NETWORK TURN-TAKING MODELS

Kornel Laskowski
Carnegie Mellon University, Pittsburgh PA, USA
Voci Technologies, Inc., Pittsburgh PA, USA

Goals

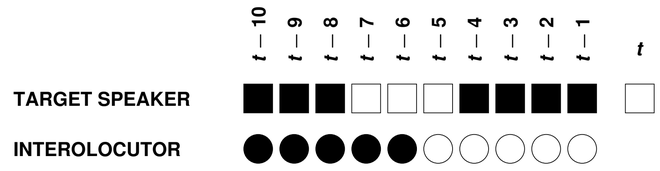
- extend **NEURAL NETWORKS**
- to handle **UNDEFINED VALUES**
- fully **AUTOMATICALLY**



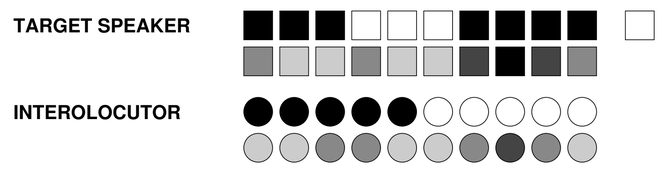
MOTIVATION: Augment stochastic turn-taking models with the ability to exploit *any* frame-synchronous features, including those which are not always defined (or modeled as attended to).

Setting

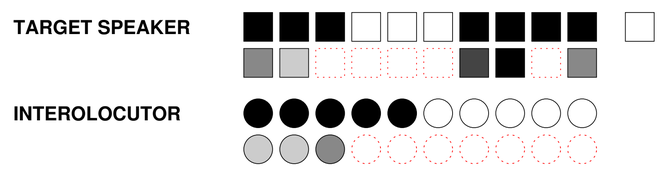
A "stochastic turn-taking model" predicts a speaker's speech activity at instant ***t*** given that speaker's and their interlocutors' speech activity at preceding instants:



Can extend to other features for the speaker and their interlocutors at preceding instants:



But what if some of those other features are occasionally undefined?



Findings

1. The proposed auto-imputing activation functions **ignore undefined features by design**.
2. "Undefined-ness" is expressed using an auxiliary **indicator feature**.
3. Prediction results using the proposed auto-imputing activation functions yield results which are **the same as those obtained using manual "brute-force" imputation**, for an easy-to-brute-force-impute feature (energy).
4. Trends observed for unseen data are identical to those observed for the data used in algorithm development, suggesting **successful generalizability**.

Standard Activation Functions

Dot-Tanh Function (DOT-TANH)

For input variable $x_i \in (-\infty, +\infty)$,

$$h_j = \tanh z_j$$
$$z_j = b_j + \sum_{i=1}^{I} w_{ji} \cdot x_i \quad .$$

(Gaussian) Radial Basis Function (GRBF)

For input variable $x_i \in (-\infty, +\infty)$,

$$h_j = \exp(z_j)$$
$$z_j = \sum_{i=1}^{I} \left(-b_{ji} \cdot (w_{ji} - x_i)^2\right) \quad .$$

Proposed Activation Functions

Bernoulli Radial Basis Function (BRBF)

For input variable $\xi_i \in \{0, 1\}$,

$$h_j = \sqrt[I]{z_j}$$
$$z_j = \prod_{i=1}^{I} (\omega_{ji})^{(\xi_i)} \cdot (1 - \omega_{ji})^{(1-\xi_i)}$$

Joint Radial Basis Function (JRBF)

For input variable $x_i \in (-\infty, +\infty)$ and **indicator** variable $\xi_i \in \{0, 1\}$,

$$h_j = \sqrt[I]{z_j}$$
$$z_j = \prod_{i=1}^{I} (\omega_{ji})^{(\xi_i)} \cdot (1 - \omega_{ji})^{(1-\xi_i)} \cdot \left(\exp\left(-b_{ji} \cdot (w_{ji} - x_i)^2\right)\right)^{(\xi_i)}$$

Impact & Future Work

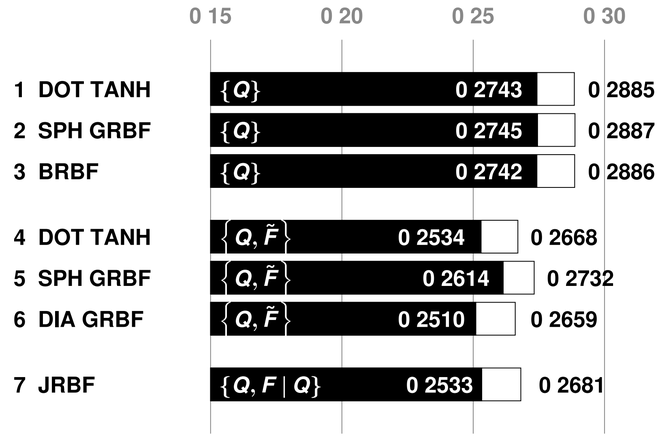
The proposed extensions:

1. Permit analysis of the conditioning of future speech activity on ***any* frame-synchronous feature with missing values** (e.g. pitch).
2. Permit empirical optimization of **normalization strategies** for such features.
3. Provide a framework for the unbiased **comparison among multiple features**, potentially conditioned on speaker identity.
4. Permit easy large-scale comparisons of feature utility **across langauges and conversation types**.
5. Permit binary-operator construction of derived indicator variables which **simulate attention span and instantaneous attentiveness to interlocutors**.

Evaluation

1. Draw training and speaker-independent test sets from the Switchboard-I Corpus (2-party).
2. Per conversation, segment speech activity ***Q*** trajectories for both parties into 100-ms frames.
3. Use 10 frames of history for prediction.
4. Per conversation and per participant, compute the prediction cross-entropy (in bits per 100 ms).
5. Pick a feature ***F*** (in addition to speech activity) which is easy to "brute-force" impute; here: ***F*** is energy.
6. Set speech activity as the indicator variable for energy.
7. Compare "brute-force"-imputed to auto-imputed cross-entropies.

DevSet Results



Note: 4, 5, and 6 are unfair comparisons to 7, since manual "brute-force" imputation was realized using per-conversation, per-participant shifted ***Z***-normalization, which uses *all* frames at every instant, including future frames.

EvalSet Results

Note: Observed trends are identical to those for DevSet.

kornel@cs.cmu.edu