Speaker confusion models

Contents

1 Models

1.1 Simple model

This simple model assumes that confusion rates (the probabilities that the algorithm attributes a vocalization from a certain speaker to another speaker) depend on the children only, and that they all derive from the same distribution, regardless of the corpus (and the surveyed population).

The simple model is defined as follows:

$$\mu_{ij} \sim \mathcal{U}(0,1) \tag{1}$$
$$\eta_{ij} \sim \text{Pareto}(1.5,1) \tag{2}$$
$$\alpha_{ij} = \mu_{ij}\eta_{ij} \tag{3}$$
$$\beta_{ij} = (1-\mu_{ij})\eta_{ij} \tag{4}$$
$$p_{c,i,j}|\alpha_{ij},\beta_{ij} \sim \text{Beta}(\alpha_{\text{ij}},\beta_{\text{ij}}) \tag{5}$$
$$N_{k,i,j}|t_{k,j},p_{c,j,i} \sim \text{Binomial}(\text{t}_{\text{k,j}},\text{p}_{\text{c,i,j}}) \tag{6}$$

Where:

- i is the speaker the diarizer returns (one of FEM, MAL, CHI, OCH)
- j is the speaker the human detected (one of FEM, MAL, CHI, OCH)
- k is a clip (i.e., a recording section that has been annotated by both a human and a diarizer)
- c is the key child to which a clip belongs

- $N_{k,i,j}$ is the number of vocalizations the human attributed to j and the diarizer attributed to i for the clip k (i and j could be the same or different categories)
- $t_{k,j}$ is the number of vocalizations returned by the human for the clip k and speaker j observed in the data
- $p_{c,i,j}$ is the probability that a vocalization from the speaker j will trigger a detection for the speaker i, for the child c.
- α_{ij} are the α hyperparameters for the Beta priors
- β_{ij} are the β hyperparameters for the Beta priors
- $\mu_{ij} = \alpha_{ij}/(\alpha_{ij} + \beta_{ij})$ are the success probabilities of the Beta priors
- $\eta_{ij} = \alpha_{ij} + \beta_{ij}$ are the effective sample sizes of the Beta priors

1.2 Model with corpus bias

We extended the previous model by including the effect of potential biases at the level of each corpus. In this model, the confusion rates do not directly derive from a Beta distribution as in (5); they are shifted by some amount depending on the corpus:

$$\sigma_{i,j} \sim \text{HalfNormal}(0, 1) \tag{7}$$
$$b_{\text{corpus},i,j} \sim \text{Normal}(0, \sigma_{i,j}) \tag{8}$$
$$\pi_{c,i,j} | \alpha_{ij}, \beta_{ij} \sim \text{Beta}(\alpha_{\text{ij}}, \beta_{\text{ij}}) \tag{9}$$
$$\text{logit}(p_{c,i,j}) = \text{logit}(\pi_{c,i,j}) + b_{\text{corpus},i,j} \tag{10}$$

In this model, $\pi_{c,i,j}$ (which still derive from a Beta distribution) captures the child-level effects, and $b_{\text{corpus},i,j}$ captures corpus-level biases.

2 Synthetic datasets

We generate datasets under the null-hypothesis, i.e. the hypothesis that the amounts of speech from each speaker are uncorrelated:

$$t_{c,j} | \lambda_{c,j} \sim \text{Poisson}(\lambda_{c,j}) \tag{11}$$
$$\lambda_{c,j} \sim \text{Gamma}(a_j, b_j) \tag{12}$$
$$c \in [\![1, n_{\text{children}}]\!] \tag{13}$$

Where:

- $t_{c,j}$ is the amount of true vocalizations from speaker j of child c

- $\lambda_{c,j}$ is the latent expected amount of vocalizations for the speaker j and child c (assuming 9 recorded hours per child)
- n_{children} is the number of children
- a_j and b_j are parameters fitted to speech rates distribution derived from manual annotations of recordings within 9am and 6pm.

We simultaneously simulate the effect of the diarization algorithm by applying the selected model to generated datasets:

$$v_{c,i,j}|t_{c,j}, p_{c,i,j} \sim \text{Binomial}(t_{c,j}, p_{c,i,j}) \tag{14}$$

$$v_{c,i} = \sum_j v_{c,i,j} \tag{15}$$

$$c \in [\![1, n_{\text{children}}]\!] \tag{16}$$

Where $p_{c,i,j}$ is sampled according to the distributions derived using the selected model ((5) or (10)).

For the model that includes corpus-level bias, the corpus from which the corresponding bias should be applied is defined by the user.