A GEOMETRIC INTERPRETATION OF NON-TARGET-NORMALIZED MAXIMUM CROSS-CHANNEL CORRELATION FOR VOCAL ACTIVITY DETECTION IN MEETINGS

Kornel Laskowski
interACT, Universität Karlsruhe, Karlsruhe, Germany
`kornel@ira.uka.de`

Tanja Schultz
interACT, Carnegie Mellon University, Pittsburgh PA, USA
`tanja@cs.cmu.edu`

PROBLEM: **WHO spoke WHEN** in headset microphone recordings of meetings.
MAIN ISSUE: Crosstalk.
SOLUTION: Train full covariance Gaussian speech/non-speech models on the ***test data.***

BUT: Need an **unsupervised initial label assignment** algorithm.

Signal at source *s(t)* has power:

$$\wp_s = \int_\Omega s^2(t)\,dt$$

Signal at microphone *k*:

$$m_k(t) = A_k\left(\frac{1}{d_k}\, s\left(t-\frac{d_k}{c}\right)+\eta_k(t)\right)$$

$$\wp_{\eta_k} = \int_\Omega \eta_k^2(t)\,dt$$

Cross-correlation between channels *j* and *k*:

$$\varphi_{jk}(\tau) = \int_\Omega m_j(t)\cdot m_k(t-\tau)\,dt = \int_\Omega \frac{A_jA_k}{d_jd_k}\, s\left(t-\frac{d_j}{c}\right) s\left(t-\frac{d_k}{c}-\tau\right)dt$$

Maximum cross-correlation:

$$\max_\tau \varphi_{jk}(\tau) = \varphi_{jk}\left(\frac{d_j-d_k}{c}\right) = \frac{A_jA_k}{d_jd_k}\int_\Omega s^2\left(t-\frac{d_j}{c}\right)dt \cong \frac{A_jA_k}{d_jd_k}\wp_s$$

To describe channel *k*, **Non-Target Normalize** the maximum cross-correlation:

$$\frac{\max_\tau \varphi_{jk}(\tau)}{\varphi_{jj}(0)} = \frac{\frac{A_jA_k}{d_jd_k}\wp_s}{A_j^2\left(\frac{1}{d_j^2}\wp_s+\wp_{\eta_j}\right)} = \frac{A_k}{A_j}\cdot\left[1-\frac{\wp_{\eta_j}}{\frac{1}{d_j^2}\wp_s+\wp_{\eta_j}}\right]\cdot\frac{d_j}{d_k} \approx \frac{d_j}{d_k}$$

When

$$\sqrt[K-1]{\prod_{j\neq k} d_j} > d_k \quad \text{then} \quad \prod_{j\neq k}\frac{d_j}{d_k} > 1$$

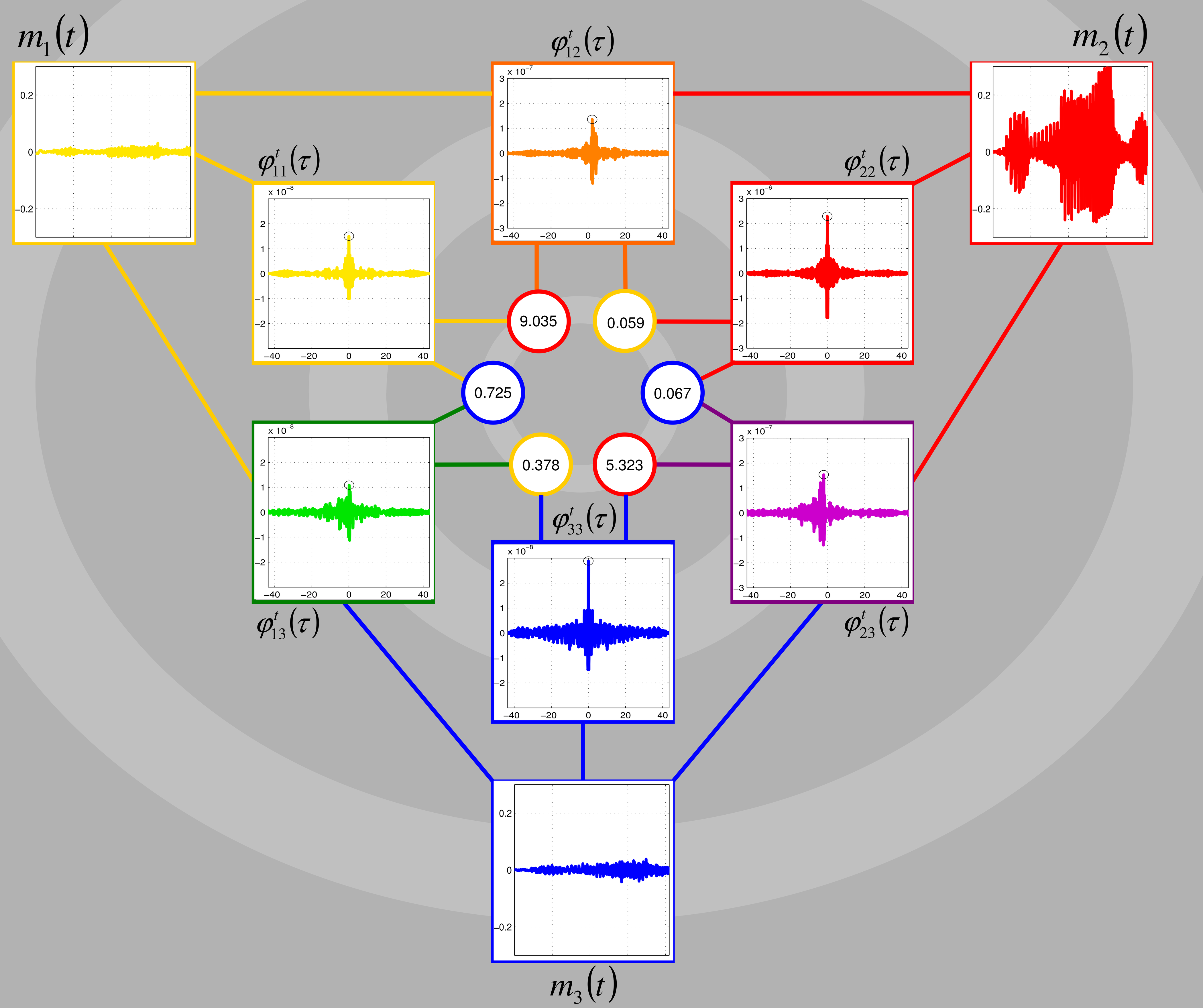
and

$$\log\prod_{j\neq k}\frac{d_j}{d_k} = \sum_{j\neq k}\log\frac{\max_\tau \varphi^t_{jk}(\tau)}{\varphi^t_{jj}(0)} > 0$$

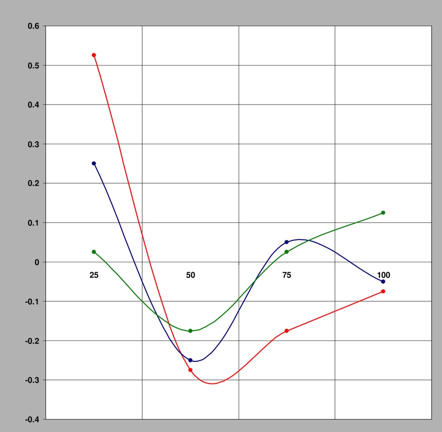
Therefore, the initial label assignment criterion

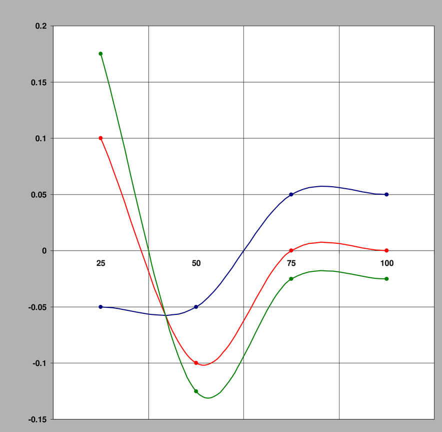
$$\mathbf{q}_t[k] = \begin{cases} speech & \text{if } \sum_{j\neq k}\log\frac{\max_\tau \varphi^t_{jk}(\tau)}{\varphi^t_{jj}(0)} > 0 \\ nonspeech & \text{otherwise}\end{cases}$$

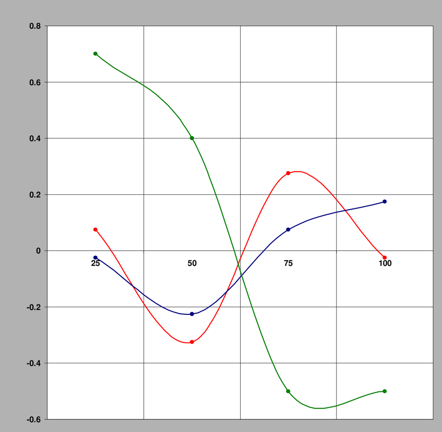
assigns speech to channel k when the distance from the source to microphone k is smaller than the geometric mean of the distances from the source to all the remaining microphones



IMPORTANT: Ω must be large enough to accommodate the true inter-speaker separation


`rt05s_eval`


`rt05s_eval*`


`rt06s_eval`