A Bayesian Network Approach Combining Pitch and Spectral Envelope Features to Reduce Channel Mismatch in Speaker Verification and Forensic Speaker Recognition

Mijail Arcienega, Anil Alexander, Philipp Zimmermann*, Andrzej Drygajlo

Interspeech 2005, Lisboa, Portugal

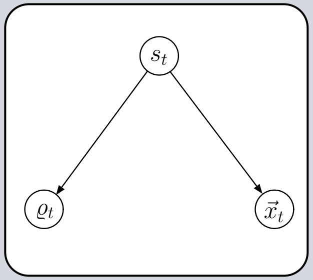
Context:
- Channel mismatch
- Speaker verification
- Forensic speaker recognition

Key Points:
- Bayesian networks
- Combining pitch and spectral envelope features

The Bayesian Network

Pitch ϱ and spectral envelope features $\vec{x}$ are conditionally independent, given the voicing status s.

$$p(\vec{x}_t|\varrho_t, s_t) = p(\vec{x}_t|s_t)$$



$$
\begin{aligned}
p(s=i) &= w_i, \\
p(\vec{x}|s=i) &\quad \text{defined by two Gaussian mixtures } \lambda_i^{\vec{x}}, \\
p(\varrho|s=1) &\quad \text{defined by one Gaussian mixture } \lambda^{\varrho}, \\
p(\varrho=0|s=2)=1 \;&;\; p(\varrho \neq 0|s=2)=0.
\end{aligned}
$$

Likelihood Estimation

Definitions:

$O = \{\eta_1, ..., \eta_T\}$ Set of testing data

where $\eta_t = \{\varrho_t, \vec{x}_t\}$

$S = \{s_1, ..., s_T\}$ Set of voicing status values

Likelihood Expression:

$$p(O|S,\lambda) = p(X|S,\lambda) \cdot p(P|S,\lambda) \;;$$

$$\Downarrow$$

$$p(O|S,\lambda) = p(X_V|\lambda_1^{\vec{x}}) \cdot p(X_U|\lambda_2^{\vec{x}}) \cdot p(P_V|\lambda^{\varrho}) \;;$$

The Database: EPFL-IPSC03

Forensic speaker recognition database (EPFL-IPSC03)

Six speech segments (15 to 180 seconds) for 60 Swiss French speakers which include recordings through:

- switched public telephone network (PSTN).
- global system for mobile communications (GSM).
- direct recording in the calling room, via a digital recorder (room).

Speaker Verification Results

Table 1: EERs when the training data is speech recorded through PSTN

Speech used for Tests	GMM-UBM EER [%]	BN system EER [%]
PSTN	4.8	3.3
GSM	42.3	31.9
Room	37.5	22.5

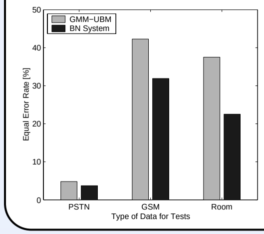
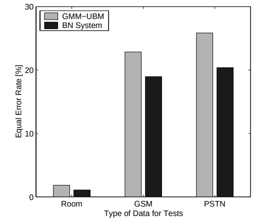
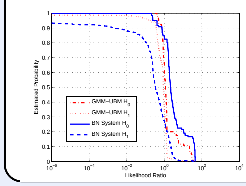


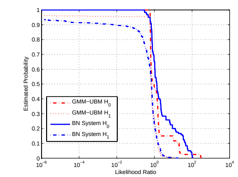
Table 1: EERs when the training data is speech recorded in the calling room.

Speech used for Tests	GMM-UBM EER [%]	BN System EER [%]
Room	1.8	1.0
GSM	22.8	18.9
PSTN	25.8	20.4



Forensic Speaker Recognition Evaluation





CONCLUSIONS:

- Convolutional modifications introduced by the PSTN or GSM channel severely affect the spectral envelope features but have almost no influence on the pitch values.
- The Bayesian network efficiently combines both features and improves both speaker verification and forensic speaker recognition systems.

Signal Processing Institute
Swiss Federal Institute of Technology, Lausanne

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

***Institut de Police Scientifique**
University of Lausanne