# BAYESIAN VOCAL TRACT MODEL ESTIMATES OF NASAL STOPS FOR SPEAKER VERIFICATION

Ewald Enzinger[1,2], Christian H. Kasess[1]
[1]Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria
[2]School of Elec. Eng. & Telecom., University of New South Wales, Sydney, Australia

Austrian Academy of Sciences

## Introduction

**Context:** Development of features for Forensic Voice Comparison

- Requires good speaker discrimination under recording condition mismatch
- Preference for more easily interpretable features

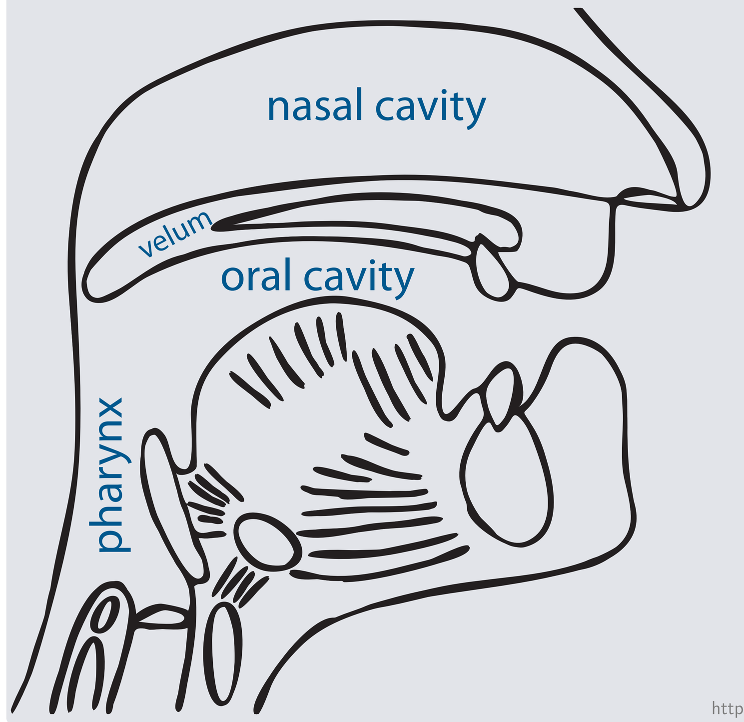
Nasal consonants are an important source of speaker-discriminating information.

- Relatively fixed structure of vocal and nasal cavity
  ➡ potentially low within-speaker variability
- Complicated structure of nasal cavity & asymmetries in paranasal cavities (sinuses)
  ➡ high between-speaker variability

**Proposed Features:** Parameters of branched-tube oral/nasal tract (VT) model of nasal consonants

**Aim of this study:** Evaluation of speaker-verification performance in controlled mismatched conditions common in forensic casework.

## The vocal tract (VT)

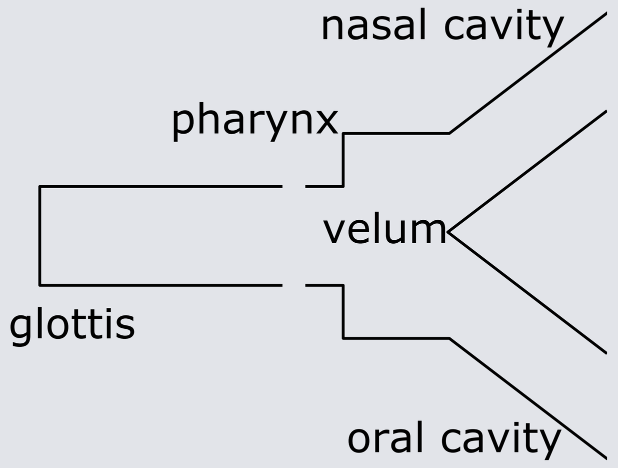



http://pegasus.cc.ucf.edu/~cnye/vocal tract pic.htm

- Roughly three cavities
  - pharyngeal
  - oral
  - nasal cavity
- Oral vowel production
  - Nasal section closed off by velum
- Nasals and nasalized vowels
  - Nasal section coupled
  - Oral section closed for nasal stops

## Branched-tube vocal tract model

Three tubes representing the pharyngeal, nasal, and oral cavity, having $L$, $M$, and $N$ segments, respectively [1, 2]. The three tubes are coupled at the velar junction. Using continuity on flow and pressure a rational transfer function $H(z) = A^{-}1(z)B(z)$ can be defined with the denominator polynomial of degree M+N+L given as:




$$A(z) = (1\ \ \mu_{M+L}) \prod_{k=M+L-1}^{M} \begin{pmatrix} 1 & \mu_k \\ \mu_k z^{-1} & z^{-1} \end{pmatrix} \begin{pmatrix} P(z) & Q(z) \\ R(z) & S(z) \end{pmatrix} \prod_{l=M-1}^{1} \begin{pmatrix} 1 & \mu_l \\ \mu_l z^{-1} & z^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ z^{-1} \end{pmatrix}$$

The forward and backward flow components $C^{\pm}(z)$ in the oral cavity are combined to give the numerator polynomial of degree $N$

$$B(z) = (1\ -1) \begin{pmatrix} C^{+}(z) \\ C^{-}(z) \end{pmatrix} = (1\ -1) \prod_{k=N-1}^{1} \begin{pmatrix} 1 & \tilde{\mu}_k \\ \tilde{\mu}_k z^{-1} & z^{-1} \end{pmatrix} \begin{pmatrix} 1 \\ \tilde{\mu}_0 z^{-1} \end{pmatrix}.$$

$\tilde{\mu}_1, \ldots, \tilde{\mu}_{N-1}$ are the reflection coefficients for the oral part, $\mu_1 \ldots, \mu_{M-1}$ are the reflection coefficients for the nasal part and $\mu_M \ldots, \mu_{M+L}$ the reflection coefficients for the pharyngeal cavity. Polynomials $P$, $Q$, $R$ and $S$ are functions of the polynomials $C^{\pm}(z)$ and the ratio of oral and nasal cross-section areas at the velum:

$$\sigma = \frac{\tilde{A}_{N-1}}{A_{M-1} + \tilde{A}_{N-1}}.$$

Thus, the model is parameterized by $\boldsymbol{\mu} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_{N-1}, \mu_1 \ldots, \mu_{M+L}, \sigma)$.

## Variational Bayes estimation scheme

The model is directly estimated from the (pre-emphasized) log-envelope $\mathbf{y}$:

$$\mathbf{y}_j = \log \bar{H}(\boldsymbol{\theta}, \boldsymbol{\omega}_j) + \boldsymbol{\epsilon}_j.$$

For the $j$-th frequency $\boldsymbol{\omega}_j$ the function $\bar{H}$ evaluates the non-linear transformation from a set of vocal tract parameters $\boldsymbol{\theta}$ to the transfer function. A sigmoidal mapping from the unrestricted $\boldsymbol{\theta}_i$ to $\mu_i$ is also included to accomodate the restricton of the reflection coeffcients to the open interval $(-1, 1)$. These parameters $\boldsymbol{\theta}$ form the VT features (VT-$\boldsymbol{\theta}$) are used in the experiments.

The Bayesian model for the estimation scheme is given as

$$p(\boldsymbol{\theta}, \tau, \boldsymbol{\Pi} | \mathbf{y}) \propto p(\mathbf{y} | \boldsymbol{\theta}, \tau)\, p(\tau)\, p(\boldsymbol{\theta} | \boldsymbol{\Pi})\, p(\boldsymbol{\Pi}),$$

where $\tau$ is the estimation error precision (i.e., the inverse variance)

$$p(\mathbf{y} | \boldsymbol{\theta}, \tau) = \mathcal{N}(\mathbf{y}; \log \bar{H}(\boldsymbol{\theta}, \boldsymbol{\omega}), \tau \mathbf{I})$$

and $\boldsymbol{\Pi}$ is the precision matrix (governed by a Gamma-hyperprior) of the smoothness prior for the vocal tract parameters

$$p(\boldsymbol{\theta} | \boldsymbol{\Pi})\, p(\boldsymbol{\Pi}) = \mathcal{N}(\boldsymbol{\theta}; \mathbf{0}, \boldsymbol{\Pi}) \prod_i \mathrm{Gam}(\boldsymbol{\Pi}_i; \mathbf{a}_i, \mathbf{b}_i).$$

In variational Bayes (VB) the posterior is factorized into a product of distributions:

$$p(\theta, \tau, \Pi | y) = q(\theta, \tau, \Pi) = q(\theta) q(\tau) q(\Pi)$$

Two assumptions about the posterior density $q(\boldsymbol{\theta}, \tau, \boldsymbol{\Pi})$ are necessary. First, $q(\boldsymbol{\theta}, \tau, \boldsymbol{\Pi})$ factors as $q(\boldsymbol{\theta}) q(\tau) q(\boldsymbol{\Pi})$. Second, as in the original scheme, $q(\boldsymbol{\theta})$ is assumed to be normal. Integrals are calculated approximately using the unscented transform [3].

## Speaker verification experiments and procedures

- /n/ tokens of 103 male adult German speakers in the Pool2010 corpus [4]
- Conditions: Normal and high vocal effort, high-quality and mobile-telephone channels
- Automatic phone-level alignment of /n/ tokens [5], followed by auditory validation
- Data was split in sets of 20/20/63 speakers for PLDA model training, development, and evaluation sets, respectively.

13 Mel-frequency cepstral coefficients (MFCCs) were extracted from the same 30 ms long portion of the tokens used for VT estimation and were used as baseline features for comparison (Hanning window, no pre-emphasis, 26 triangular filters with 50% overlap).

## PLDA modeling and Likelihood Ratio calculation

VT parameters (VT-$\boldsymbol{\theta}$) as well as MFCCs were modeled using probabilistic linear discriminant analysis [6]. Feature vectors are assumed to be generated by a generative model:

$$x_{ij} = \mu + \mathbf{F}\mathbf{h}_i + \mathbf{G}\mathbf{w}_{ij} + \epsilon_{ij}.$$

$x_{ij}$ denotes the $j$th observation (VT-$\boldsymbol{\theta}$s or MFCCs) of speaker $i$, $\mu + \mathbf{F}\mathbf{h}_i$ describes the between-speaker variability, and $\mathbf{G}\mathbf{w}_{ij} + \epsilon_{ij}$ the within-speaker variability. As in [6] we use a Gaussian residual term $\epsilon_{ij}$ with diagonal covariance $\boldsymbol{\Sigma}$. The priors of the latent variables $\mathbf{h}_i$ and $\mathbf{w}_{ij}$ are assumed to be Gaussian.

Given mean vectors $\bar{x}_1$ and $\bar{x}_2$ obtained from observations of /n/ tokens in the training (enrollment) and test portions of a verification trial, a score $s$ is calculated as a likelihood ratio with respect to two hypotheses, that both vectors share the same latent identity variable ($H_1$), or that they were generated from different latent identity variables ($H_2$):

$$s = \frac{p(\bar{x}_1, \bar{x}_2 | H_1)}{p(\bar{x}_1 | H_2) p(\bar{x}_2 | H_2)}.$$

Logistic regression was used for calibration [7] and to fuse the scores from VT-$\boldsymbol{\theta}$ and MFCC based systems [8]. Its parameters were trained on scores obtained from tests on the development set.

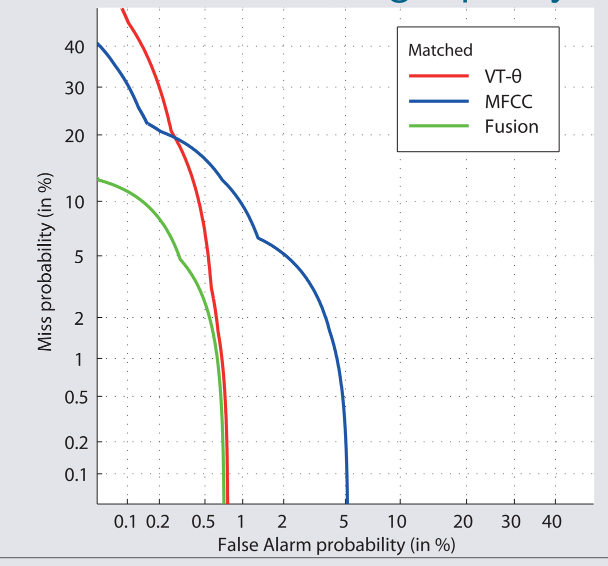
## Results

### Vocal tract prior settings:

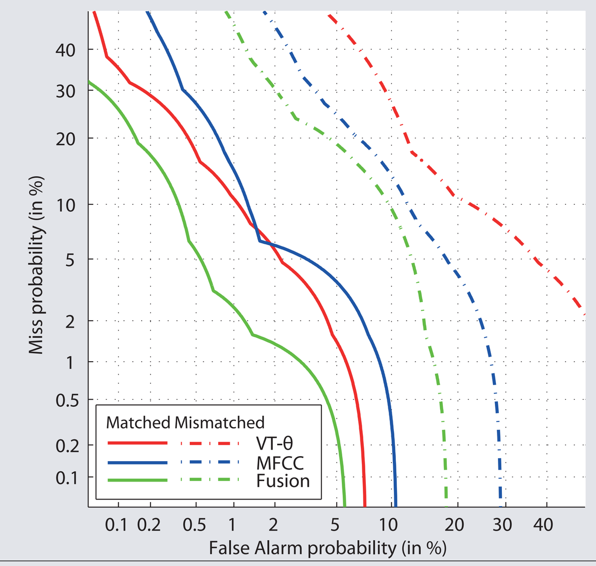
Six different VT prior settings for the $\mathbf{a}_i$ (10, 20, 50, 100, 100, 200) and $\mathbf{b}_i$ (1, 1, 1, 2, 1, 2) were evaluated on the development set. The expected value for precision is given as $a/b$. Results suggest that higher values for the precision lead to better speaker verification performance.

| $\mathbf{a}_i/b_i$ | 10/1 | 20/1 | 50/1 | 100/2 | 100/1 | 200/2 |
|---|---|---|---|---|---|---|
| EER (%) | 2.90 | 3.26 | 3.62 | 4.00 | 2.78 | 1.52 |
| $C_{llr}$ | 0.180 | 0.185 | 0.179 | 0.181 | 0.108 | 0.082 |

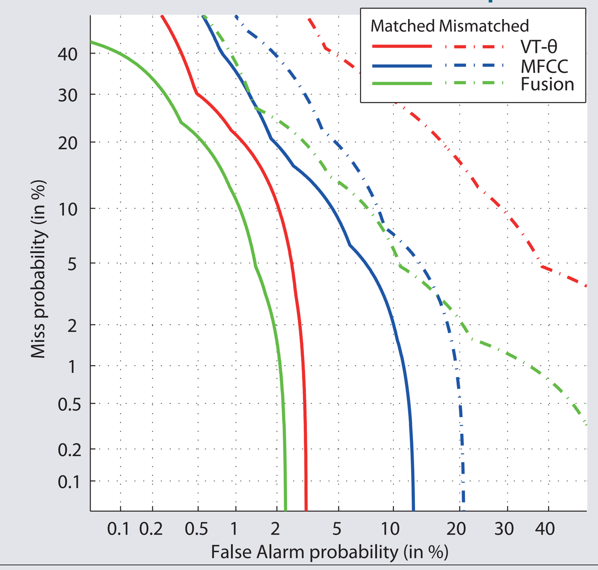
### Normal vocal effort, high-quality recs.




### High vocal effort, high-quality recs.




### Normal vocal effort, mobile-telephone recs.




| | Normal vocal effort, high-quality recordings | | High vocal effort high-quality recordings | | | | Normal vocal effort, mobile-telephone channel | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Matched | | Matched | | Mismatched | | Matched | | Mismatched | |
| | **EER** | $\mathbf{C_{llr}}$ | **EER** | $\mathbf{C_{llr}}$ | **EER** | $\mathbf{C_{llr}}$ | **EER** | $\mathbf{C_{llr}}$ | **EER** | $\mathbf{C_{llr}}$ |
| VT-$\boldsymbol{\theta}$ | 0.7 | 0.055 | 3.30 | 0.317 | 15.00 | 0.574 | 2.70 | 0.155 | 18.30 | 1.265 |
| MFCC | 3.1 | 0.158 | 4.20 | 0.201 | 11.30 | 0.405 | 6.10 | 0.229 | 8.80 | 0.401 |
| Fusion | 0.7 | 0.036 | 1.50 | 0.133 | 9.80 | 0.333 | 1.90 | 0.102 | 8.40 | 0.686 |

## Discussion and Conclusion

This study assesses the performance of physiologically motivated vocal tract model estimates of alveolar nasal stop (/n/) tokens in speaker verification experiments.

The main conclusions are:

- Performance increased with higher precision values in the Bayesian VT estimation.
- Performance of VT-$\boldsymbol{\theta}$ based systems compared favorably to that of MFCC based systems under matched conditions, but not under mismatched recording conditions.
- Fusion of both systems generally improved upon both individual systems, indicating that they offer complementary information.

Possible causes for lack of robustness:

- Differences in fundamental frequency induced by high vocal effort [4] may have a profound effect on spectral envelope estimate, leading to different VT model estimates.
- Adaptive Multi-Rate (AMR) codec used in GSM and UMTS mobile telephone networks uses order 10 linear prediction to encode the spectral envelope, which may affect the vocal tract estimation.

## References

1 K. Schnell, PhD thesis, Goethe University Frankfurt am Main, 2003.
2 I.-T. Lim and B. Lee, *IEEE TASP*, 4(2), 81–88, 1996.
3 S. Julier and J. Uhlmann, *Proc. SPIE Conf. on Sig. Proc., Sensor Fusion, and Target Recogn.*, vol. 3068, pp. 182–193, 1997.
4 M. Jessen, O. Köster, and S. Gfroerer, *Int. J. Speech, Language, and the Law*, 12(2), 174–213, 2005.
5 S. Rapp, *Proc. ELSNET Goes East and IMACS Workshop*, 1995.
6 S. J. D. Prince and J. H. Elder, *IEEE 11th International Conference on Computer Vision (ICCV)*, pp. 1–8, 2007.
7 N. Brümmer and J. du Preez, *Computer Speech and Language*, 20, 230–275, 2006.
8 S. Pigeon, P. Druyts, and P. Verlinde, *Digital Signal Process.*, 10, 237–248, 2000.

For further questions please contact ewald.enzinger@oeaw.ac.at